KNOWLEDGE DISCOVERY THROUGH ONTOLOGY-BASED KNOWLEDGE MANAGEMENT SYSTEMS FOR DIGITAL LIBRARIES: A CONCEPTUAL FRAMEWORK

M. A. Angrosh

Abstract

In recent times, Knowledge Organization Systems (KOSs) such as ontologies have emerged as promising developments in the areas of knowledge representation, natural language processing, information retrieval, knowledge discovery and semantic web activities. The primary objective of these developments has been oriented towards developing effective and efficient information-based knowledge management systems. The ability of ontologies in combining together two different tasks of 'information retrieval' and 'knowledge discovery', bestows a prominent place for these tools in digital libraries. The present paper while tracing the importance of ontologies in the context of digital libraries, researches upon the possibility of utilizing readily available 'subject classification systems' and 'subject headings' in conjunction with subject-specific domain ontologies for developing effective ontologybased knowledge management systems for knowledge sharing in digital libraries. Accordingly, the paper while establishing the place of ontologies in digital libraries, draws upon an OWL (web ontology language) illustration to explain the concept of information mapping in ontology-based information systems. Further, the paper while illustrating the operations of knowledge agents, identifies some open sources which can be harped upon for developing semantic-web applications.

Keywords : Ontology; Ontology-based Information Systems; Digital Libraries; Web Ontology Language (OWL); Knowledge Agents.

1. Background

Strategic In recent years, libraries and information systems have witnessed a major turnaround by budging towards digital medium from traditional print medium. Albeit print content still form a major component of libraries and information centers, the march towards digital medium is fast and rapid. The notion of 'Digital Libraries', which was looked upon in skepticism a few years ago, has today become a simple reality. During this transformation process, while major focus was laid upon digital content production, collection and storage, the area of digital content retrieval received very little attention. This has resulted in serious problems with respect to information retrieval in digital form. The difficulty faced in retrieving information on the Internet medium is a classic example of this phenomenon. Efforts in the direction of developing effective information retrieval mechanisms in digital form have resulted in the evolution of Knowledge Organization Systems (KOS). KOSs are now being observed as vital digital information management tools. Ontologies are one such form of knowledge organization systems, which have emerged as promising developments in the areas of knowledge representation, natural language processing, information retrieval, knowledge discovery and semantic web activities. The ability of ontologies in combining together two different tasks of 'information retrieval' and 'knowledge discovery', bestows a prominent place for these tools in digital libraries. The present paper while tracing the importance of ontologies in the context of digital libraries, researches upon the possibility of utilizing readily available 'subject classification systems' and 'subject headings' in conjunction with subjectspecific domain ontologies for developing ontology-based knowledge management systems for

knowledge sharing in digital libraries. Accordingly, the paper while outlining the conceptual framework of ontology-based information systems for digital libraries, draws upon the functioning of 'Knowledge Agents' by using an illustrative agri-pest ontology developed using web ontology language (OWL). The paper while drawing upon the operations of ontology-based knowledge agents, identifies few open source semantic tools that can be utilized for developing semantic-web applications, particularly in the context of digital libraries.

2. Ontologies: What are they?

The term 'ontology', coined as early as seventeenth century, fundamentally belong to the field of philosophy, owing its essence to 'metaphysics', primarily concerned with 'enquiry into existence of being' (Heidegger 1999). The field of Artificial Intelligence employ the term to purport the science of specification of existing concepts. Ontologies are defined as 'explicit specification of conceptualization', wherein 'conceptualization' implies couching of knowledge of a particular domain (or general) in terms of entities (things, relations and constraints) and 'specification' refers to the representation of conceptualization in concrete form (Gruber 1993). Ontologies are observed as a formal explicit description of (a) concepts in a domain of discourse; (b) properties of each concept describing various features and attributes of the concepts and (c) restrictions on slots (Noy and McGuinness, 2001). In principality, besides identifying important classes in a domain under study, an ontology lay down stipulations for organizing these classes. While the structure of an ontology is observed to be simplistic resembling a taxonomy, the existence of inference and deduction rules, reasoning and classification services lends an air of supremacy to these systems.

2.1 Ontologies: Do we need them?

The current practice of information search is too simplistic with the process involving keying in relative terms, which results in a listing of zillions of records, leading to a confused state of users, characterized by their helplessness of not getting what they wants. While free-text word-based search engines are observed to return innumerable irrelevant hits, full-text search methods are found to be incapable of processing structured queries. Further, current search methods are also observed to yield a low precision and recall rate. An interesting study carried out by Feldman and Sherman (2001) for establishing the cost of not finding digital information, observes that an enterprise employing 1000 knowledge workers would waste approximately \$ 2.5 to \$ 3.5 million per year in searching non-existing information. The study while underlining the importance of information in decision-making, asserts precise and perfect information as major assets of organizations. The exorbitant growth of information has further complicated the scenario. The amount of digital information available online today, as estimated by IBM is of the order of one exabyte (Liautaud, B. and Hammond, M., 2000).

Such problems currently faced with present search strategies can be alleviated if systems are designed which no longer search for matching words but work on semantic mechanism primarily built around information and their relationships that are present in web pages / documents. The semantic mechanisms focus on evolving a system where information meanings are represented in a manner understandable to both machines as well as humans. Ontology-driven systems are observed to play a vital role in such initiatives.

Various reasons, which strongly advocate the use of ontologies in designing information systems, include:

- facilitate achievement of a higher precision and recall rate;
- facilitate representation of complex relationships among objects including rules and axioms missing from semantic networks;

- facilitate machine-to-machine communication i.e., support terminological reasoning;
- ability to handle voluminous information with less resource consumption;
- effectively overcome problems arising from linguistics;
- facilitate information retrieval from diverse, heterogeneous sources, supporting both structured and unstructured data.

2.2 Ontologies: Applications

Ontologies are being effectively employed in improving existing web applications in various intensities. Heflin et al (2002) identifies the following ontology-based cases designed to provide valuable services:

- Web Portals: Web portals are web sites that collect information on a specific topic. An ontology based web portal could provide more intelligent syndication, which a conventional subject index may not be able to provide to the community. OntoWeb is a classic example of ontology-based web portal.
- Multimedia Collections: Ontologies are observed to provide semantic annotations for collections
 of images, audio and other non-textual objects. Such annotations are capable of supporting
 indexing and search. This characteristic can be effectively utilized in designing innovative systems
 for digital libraries.
- Corporate Web-Site Management: Ontologies can be effectively utilized for indexing corporate
 web documents and provide better services. Such systems can serve as a point of access to
 different individuals serving in different capacities, searching information from different angles.
- Design Documentation: Design documentation relates to a large body of engineering documentation such as aerospace industry. These documents sets are observed to have a hierarchical structure, which differ between sets. An information model, which facilitates exploration of information space in terms of representing items, associations between terms, properties of the items and links to documentation, which describes and defines them can be effectively built by using ontologies.
- Intelligent Agents: Intelligent agents are systems that are capable of understanding and integrating diverse information resources. Domain ontologies with terms and representative relationship between them pertinent to the specific domain under study constitute building blocks of intelligent agents. The Wine Agent is a good example of intelligent agents.
- Ubiquitous Computing: Ubiquitous computing is a newly emerging area in personal computing, characterized by the shift from dedicated computer machinery to pervasive computing capabilities embedded in our everyday activities. As the systems operability can be modeled as web services, the requirements of these systems are in line with ontological principles. Ontologies are foreseen for playing a central role in these areas.

3. Ontologies : Applications in Digital Libraries

Systems developed on ontological principles are observed to play an active role in digital libraries. The potential role played by ontologies in information management, as noted in examples above, creates new vistas for effective utilization of ontologies in digital libraries, which are primarily repositories of

electronic information. However, presently the metadata collections of electronic resources available in the digital library serve as points of access to these valuable resources. The current practice of information retrieval in digital libraries include search through special-purpose web portals specifically designed for the respective digital library, which acts as information gateway to the available contents in the digital library. Such portals acting as an interface between the user and the database, performs the tedious task of searching, locating and presenting the required data to the user in acceptable form. However these search strategies suffer from serious limitations. Non-retrieval of required data, even is present in the database can take place for various reasons viz., (a) usage of synonyms by stored documents (b) spelling mistakes (c) linguistics problems (d) inefficiency of search engines to process html documents and others in the main.

In order to overcome these inefficiencies and increase the accuracy of data retrieved, ontologies can be effectively utilized in digital information management systems. The presence of ontologies in such systems facilitates in developing an ontology layer that would agglomerate the metadata collection and subject relationships through appropriate specifications. This layer would then form the point of access to contents present in the digital library. Through usage of appropriate reasoner, stored data can be queried for obtaining more relative and accurate information. A schematic diagram portraying the ontology layer as points of access, integrating metadata of different kinds of electronic resources available in the library is depicted in *Figure 1.*(Footnotes)

Electronic Metadata Journals/ Subject Otology Electronic Metadata Geographic Information Otology Retrieval Multimedia Metadata Collections Information Ontology ONTOLOGY Internet **REASONER** Resources Metadata **DIGITAL ONTOLOGY LIBRARY LAYER**

Figure 1
Ontology Layer Integrating Metadata of Electronic Resources in Digital Library

3.1 Ontology Modeling for Digital Libraries

An ontology layer connecting metadata resources can provide the required functionality in developing effective information management systems. The process of ontology building involves execution of a chain of activities primarily aimed at developing knowledge management systems. Following Beck and Pinto (2002), the principal stages of ontology development include specification, conceptualization, formalization, implementation and maintenance. Functions carried out in the 'specification' stage in building the ontology, primarily involve identification of purpose and scope of the ontology. A feasibility study analyzing the purpose and scope of ontologies for digital libraries under focus can provide the required output for the specification stage. Typically as Noy and McGuiness (2001) suggests, the following questions needs to be addressed in this stage:

- What is the domain that the ontology will cover?
- For what purposes would the ontology be used?
- For what types of questions, the information in the ontology should provide the answer?
- Who will use and maintain the ontology?

Further, studying the feasibility of using existing ontologies is also recommended since building of ontology is observed to be a cumbersome and costly affair. The second stage of ontology building viz., 'conceptualization', fundamentally deals with description of the conceptual model in accordance with the stipulations laid down in the previous stage of 'specification'. The primary focus of this stage is to concentrate on enumeration of important terms to included in ontology. Digital libraries can make the best use of readily available enumerative subject classification systems for realizing this objective. Besides providing standardized terms for developing the ontology, classification systems can also provide a clear structure for building the ontology. Further, the intricate relationships weaving the subject matter in subject classification systems can ease the job of domain experts to a great extent in developing the ontology. Furthermore, usage of classification systems would also result in sufficient understanding of ontological systems and structure for 'information managers' managing the digital library system, which would pave ways for providing efficient information services. The next stage, i.e., 'formalization' mainly deals with transforming the conceptual description into a formal form i.e., fine-tuning the conceptual structure developed for the ontology.

The process of 'implementation' involves implementing the formalized ontology in formal knowledge representation languages. Appropriate knowledge representation language can be employed for conceptualizing formal descriptions of concepts in ontology. Knowledge representation languages facilitate symbolic representation of domain knowledge with formal standards in order to utilize these representations for making inferences and drawing conclusions. The languages that can be banked upon are Web Ontology Language (OWL), Ontology Inference Layer (OIL), DARPA Agent Markup Language (DAML), Resource Description Language (RDF), Resource Description Language Schema (RDFS) etc. Knowledge representation languages can be effectively used for defining the complex relationships among concepts and its subclasses for developing a working ontology. Finally, the 'maintenance' stage is mainly concerned with updating and correcting the implemented ontology. The framework of an ontology-based knowledge management system involving subject classification ontology and subject headings for digital libraries is shown in Figure 2.

Specification Ν Purpose, scope and W requirements identification D G E Conceptualization / Formalization W 0 R Conceptual mode & K E transformation info formal form R ONTOLOGY MANAGER (User interface) Implementation Updataion and correction of Subject Classification Subject Headings ontology Ontology Ontology Maintenance Ι Ε Ν Ν D Knowledge base / 0 Knowledge Sharing Metadata R M U S Ε 0 N Output Interface Reasoner / Query / Μ Filter and combination A N Α (User interface) G Ε R

Figure 2
Framework of Ontology-based Knowledge Management Systems for digital libraries

As illustrated in *Figure 2*, ontologies can be effectively modeled for developing knowledge management systems in digital libraries for providing efficient information services. 'Knowledge sharing system', the user interface allows an user for specifying meaningful phrases, keywords, author names and title terms

for executing a particular search. Keywords form prominent means of input for ontology-based information systems. However, author names as well as title terms can also be transformed as prospective inputs to such systems. Flow-chart discussed in the following sections of the present paper encapsulates this idea. Irrespective of the search type, the initial stages involved in processing the provided query would primarily lay upon arriving at the "key-term" defined in the domain ontology. Subject-headings ontology employed in the system facilitates in achieving this objective. These tools provide the necessary functionality through provision of correction techniques and options for entering sub-headings in the search term / phrase. Further, necessary search modifiers can be utilized for pruning the search phrase to the desired level. The schematic representation of subject classification and subject headings ontology in a knowledge representation language would be as follows:

The class "starch crops" is identified as a subclass of "fieldcrops" in subject headings ontology of agricultural crops. The obtained keywords would then become input to the subject classification (domain) ontology. The domain ontology primarily consists of concepts, attributes and instances of domain-specific knowledge. The primary objective of the domain ontology is to achieve the objective of semantic match when searching the knowledge base.

Consider an example of crop-pest domain ontology. The ontology underlying crop-pest domain would contain hierarchies and descriptions of crop pests, crop diseases and pesticide categories, along with restrictions on how particular instances can be paired together. The following form a snippet of various categories that can be developed as ontologies for deriving a crop-pest domain ontology:

Pest Category	Crop Diseases Category	Pesticide Category
- Pest Name	- Crop Diseases (damage)	- Pesticide Names
- Common Name	- Respective crop	- Chemicals
- Family Name	- Geographical area	- Properties
	- Chemical Treatment	
	- Chemical Properties	

The schematic representation of these categories in Web Ontology Language (OWL), a W3C Recommendation for explicitly representing the meaning of terms in vocabularies and relationships between those terms is shown below:

a. Pesticide Category

```
<rdfs:Class rdf:ID="Pesticide">
    <rdfs:subClassOf rdf:resource="#Chemicals/>
    <rdfs:subClassOf>
        <owl:Restriction>
        <owl:onProperty rdf:resource="#Maker/>
        <owl:minCardinality>
        1
        </owl:minCardinality>
```

According to the above specification, a 'pesticide' is a 'chemical' produced by at least one 'Maker' of type 'ChemicalFactory' from at least one chemical. Such chemicals are restricted to Pest Chemicals elsewhere in the ontology More importantly, such pesticides have different 'Chemical' properties.

b. Crop-Disease Category

Ontology definitions in a crop-disease category underlie pairing of a particular crop with a particular disease (damage). Each disease treatment aims at targeting a particular solvent pairing individual crops with respective diseases. When a user selects a kind of disease or a crop, which gets mapped to a disease, the agent would consult the "disease" definitions for restrictions on constituents of "crop" and "pesticides". All such "pesticides" types maps back to this concept.

Suppose the user has selected a particular disease "foliage damage" in "herbaceous" crop. The concept of such a damage is defined elsewhere in the ontology, but most important here is the notion of "foliage damage", which is defined as a "disease" present in herbaceous crops, which is caused by a pest known as "Acheta spp.". Furthermore, the ontology also defines that such pests can be controlled by chemical constituents having 99%g-BHC.

```
<rdfs:Class rdf:ID="FoliageDamage">
    <owl:intersectionOf rdf:parse Type="owl:collection">
        <rdfs:Class rdf:about="HerbaceousCrops"/>
        <owl:Restriction>
        <owl:onProperty rdf:resource="Acheta-spp."/>
              <owl:toClass rdf:resource="#FoliageDamage">
              </owl:Restricition>
        </owl:ntersectionOf>
              <rdfs:subClassOf
rdf:resource="#Chemical-has-99%g-Isomers-Restriction/>
</rdfs:Class</pre>
```

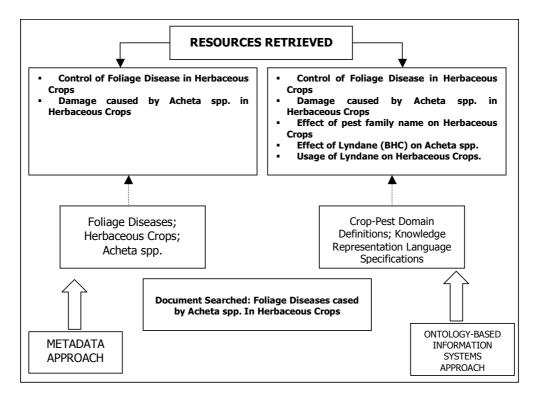
c. Pesticides Category

Coming back to the pesticides ontology, the definitions observe one chemical that would match the above restrictions is Lyndane, subset of BHC and exists as five isomers with g isomers as the active ingredient. Further, this inidividual solvent is simply defined as "Lindane", whose maker is "Agro Industries-BHC". Together with other statements in the ontology, this allows the reasoner to deduce many additional facts such as Lyndane contains 99 % of BHC.

For instance the concept of "Lyndane" specifies that all such solvents are less harmful to bees and livestock, which contain 99 % g isomers. Further, Lyndane is also a particular subset of Benzyl Hexachloride, distinguished as existing as five isomers in technical form but the active ingredient is the g isomers. It is through this additional subclass relationship that additional properties of Lyndane viz., dehydrochlorinatable by alkalis is defined elsewhere.

Design of ontology-based systems including components described above result in the development of 'Knowledge Agents'. Knowledge agents are observed go beyond web portals in retrieving the required information. Besides enabling retrieval of precise information, these systems largely facilitate in 'knowledge discovery'. The process carried out by a 'knowledge agent' in knowledge discovery can be is illustrated in *Figure 3*.

Figure 3
An Illustration of Information Retrieval by Knowledge Agent



As illustrated in Figure 3, ontology-based 'knowledge agents' function beyond matching keywords for retrieving more precise information. Irrespective of available metadata terms, knowledge agents coordinate towards retrieval of relevant information. Ontologies add value to information management systems by discharging the role of an intelligent system through interpretation of the keyed-in data, derivation of essential meaning and retrieval of appropriate information. Functions of knowledge agent can be more clearly understood from the flowchart illustrated in *Figure 4*, which draw upon various operations performed by an ontology-based information system. As seen in the flowchart, subject classification ontologies form the primary screening stage for ascertaining the existence of 'search-terms' in the domain (subject-specific) ontologies. However, it needs to be observed to build domain ontologies in relation with subject-classification ontologies. This would result in more efficient form of information retrieval from the system. The domain specific ontologies are fundamentally designed to capture knowledge about the particular discipline in question. Input terms from subject classification ontologies, defined explicitly in domain ontologies would then result in more precise and relative output.

User-Defined
Title Term
Check for the Term's
Existence in Subject
Ontology

Prompt
User for defined
Ontology

The Deciment (a) Search
on Subject Term Definitions

Check for titles with
similar author

Check for subscience in
subject
similar author

Check for titles with
similar author

Check for t

Figure 4
Flow Chart Depicting Knowledge Agent's Search Execution

More specifically, as shown in the flowchart in *Figure 4*, an ontology-based information system can be designed for searching information from three different angles viz., subject phrase, title term and author term. The ensuing section explains the various operations performed with respect to each category:

a. Subject Phrase

A user may key-in a relative subject phrase describing the information, which he is looking for. The following form the various operations executed for retrieving information with respect to subject-phrase search:

- Identification of important terms in the phrase analyzing the grammar present in the statement.
- Decision-making action checking whether the term exists in subject ontology defined for the system.

Results in either Yes or No

- If No –
- Prompt user for different phrase / statement
 - If Yes -
- Check for subject definitions in the subject ontology
- Retrieval of appropriate document in relation to definitions found in subject ontology.
- Check for definitions of identified document's definition.
- Retrieval of document (s) in relation to definitions of the identified document.
- Display to the user.

b. Title Term

A user may also key-in a term which is a part of the title. The following form the various operations executed for retrieving information with respect to title term search:

• Decision making action for term's exists in subject ontology

Results in either Yes or No

- If No –
- Prompt user for different a term.
 - If Yes -
- Check for subject definitions in the subject ontology
- Retrieval of appropriate document in relation to definitions found in subject ontology.
- Check for definitions of identified document's definition.
- Retrieval of document (s) in relation to definitions of the identified document.
- Display to the user.

(c) Author Name

A user may also key-in an author name in order to retrieve information. The following form the various operations executed for retrieving information with respect to author search:

Decision making action for titles existing with similar author.

Results in either Yes or No

- If No –
- Prompt user for different author name
 - If Yes -
- Check for Documents key word metadata
- Check for key word's existence in subject ontology

Results in either Yes or No

- If No -
- Display retrieved documents
 - If Yes –
- Check for term definitions in the subject ontology
- Retrieval of appropriate document in relation to definitions found in subject ontology
- Check for definitions of identified document's definition.
- Retrieval of document (s) in relation to definitions of the identified document.
- Display to the user.

4. Ontology-based Information System for Digital Libraries: Technical Contours

With rapid developments and increasingly easier access to IT developments with open sources round the corner, ontology-based information systems in the context digital libraries can be developed with minimum efforts. *Table 1* outlines some of the open sources, which can be effectively utilized for technically realizing an ontology-based web application in the context of digital libraries, which can then be uploaded over the Web for providing value-added services to the end user community.

An Outline of Technical Details for Developing Ontology-based Information System for Digital Libraries

SI. No.	Task	Available Open Source Tools	Result
1.	Ontology Development	Protégé	Ontology in the desired web language
		SemWeb	(RDF, OWL, DAML etc)
		OntoEdit	
2.	Database Development	Database Servers including MySQL, PERL	RDBMS Database / Knowledge base of available data
3.	Relating Database through developed Ontology	Java 1.4 and higher JDBC Connectivity Jena 2.1	Conversion of available RDBMS Data into RDF Data
4.	RDF Reasoning	Pellet; OWL Reasoner; Fact++	Performing Reasoning techniques on RDF data
5.	RDF Data Storage and Querying	Java Sesame SPARQL	Information retrieval from the RDF Data

Utilization of Java Servlet Pages (JSP) and HTML for designing interactive user interfaces will lead to the development of semantic-web based applications in the context of digital libraries aimed at providing high quality information services.

5. Summing Up: Ontologies in Digital Libraries-Future Directions

The current era, characterized by a high degree of information explosion and growth, needs appropriate information management systems in place for fully exploiting available information resources. Ontology-based information systems provide ample ways for achieving such objectives. Besides playing an important role in information retrieval, ontological principles can play a crucial role in various other areas / services of digital library. For example, ontologies can be effectively utilized for combining different data sources in order to create valuable knowledge from such combinations. For instance, data documented in different departments of the library for various specific reasons can be combined for deriving meaningful knowledge from such agglomerations. Further, ontologies can also play a major role in building library networks through effectively intertwining information resources spread over a wide geographic area. Such ontology-based library networks, besides provisioning valuable information services, would create sufficient conditions for optimum utilization of available information resources. Ontologies can also be made use for providing value added subject dissemination services and current awareness services for its users. Further, with ontologies being used for providing personalized information services for the library community can facilitate in bringing together the community's research, scholarly, development and other activities under one arena.

¹ Terabytes, petabytes, exabytes are similar to gigabytes, megabytes and kilobytes, which measure computer memory storage. A kb is 1,000 bytes, a mb 1000,000 bytes or 1,000 kilobytes. Equivalently, a gigabyte is 1,000 megabytes, a terabyte is 1,000 gigabytes, a petabyte is 1,000 terabytes and an exabyte is 1,000 petabytes.

6. References

- 1. Beck, H. and Pinto, H.S. "Overview of Approach, Methodologies, Standards and Tools for Ontologies (Ontology for Agricultural Domain)". UN: The Agricultural Ontology Service, FAO, 2002.
- 2. Boulos, KMN; Roudsari AV and Carson ER. "Towards a Semantic Medical Web: HealthCyberMap's Dublin Core Ontology in Protégé-2000", Presented at the Fifth International Protégé Workshop, Sowerby Centre for Health Informatics at Newcastle (SCHIN), University of Newcastle, England, 18th July 2001.
- 3. Buckingham, S.S.; Motta, E. and Domingue, J. "ScholOnto: An Ontology-Based Digital Library Server for Research Documents and Discourse, International Journal on Digital Libraries", 3(3), August/Sept., 2000, Springer-Verlag. (available online)
- 4. Feldman, S. and Sherman, C. "The High Cost of Not Finding Information: An IDC White Paper". Farmingham: IDC, 2001. (available online)
- 5. Gruber, Thomas R. "A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition". 5: 199-220., 1993. (available online)
- 6. Heidegger, Martin. "An Introduction to Metaphysics". New Delhi: Motilal Banarasidas Publishers, 1999.
- 7. Dennis S. Hill. "Agricultural Insect Pests of the Tropics and their Control". New York: Cambridge University Press, 1993.