Development of Indian Agricultural Research Ontology: Semantic Rich Relations Based Information Retrieval System for Vidyanidhi Digital Library

M.A. Angrosh and Shalini R. Urs

International School of Information Management
University of Mysore, Manasagangotri Mysore, India
angrosh@isim.ac.in, shalini@isim.ac.in

Abstract. Digital Libraries represent semantically rich collections of digital documents. Ontology-based information retrieval systems capture semantic relations for providing value added information services. Deviating from the regular approach of developing ontologies on the basis of domain knowledge, the present paper puts forward a novel method for developing ontologies from the semantic information available in the titles of digital documents. Such an approach gathers significance due to its simplicity in ontology development process. To examine the same, the study considered the case of Agricultural Electronic Theses and Dissertations (ETDs) present in Vidyanidhi Digital Library. The study resulted in the development of Indian Agricultural Research domain ontology, which was used for developing ontology-based information retrieval system. This paper while describing the methodology followed for developing the ontology presents the technical details of the developed system.

Keywords: Indian Agricultural Research Ontology, Ontology, Web Ontology Language, Semantic Web, Vidyanidhi Digital Library.

1 Introduction

The field of Information Retrieval is a central area of research in Digital Libraries. Information Retrieval (IR) is a process of finding all relevant documents from a document collection, satisfying user information need [1]. Unfortunately, currently employed information retrieval mechanisms suffer from various limitations. Issues such as information overload, rapid technological developments, fluctuating user trends and behaviour call for better IR mechanisms. The emerging Semantic Web technologies such as ontologies promise knowledge-based systems capable of performing crucial tasks of information retrieval and extraction [2]. Domain ontology based systems supporting navigation and querying facilities form ideal information retrieval systems for digital libraries. The reasoning and querying capabilities offer valuable search strategies for digital libraries.

Ontology-based IR systems mainly rely on domain ontologies, which are further extended for information retrieval. Development of domain ontologies is not only costly [3], time consuming and cumbersome, but also leads to difficulties particularly

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 400–409, 2007.

at the time of mapping document instances to the ontology. Further, the high volume of knowledge represented in the ontology may not be used in its entirety. Thus, instead of representing the entire knowledge into domain ontology and mapping documents to the ontology, it would be appropriate to develop ontology, based on the available documents and deploy the same for information retrieval. This would also facilitate in the easy mapping of documents to the ontology. The present paper puts forward a simple yet powerful method for developing ontologies from the information present in the titles of electronic documents, resulting in effective information retrieval systems for Digital Libraries. This paper is the result of the study carried at Vidyanidhi Digital Library (VDL). The study focused on developing ontology from the information available in the titles of Agricultural Electronic Theses and Dissertations (ETDs) present in VDL, resulting in 'Indian Agricultural Research' ontology. This was mapped to Agricultural ETDs for developing a knowledge base, which was used for information retrieval.

The paper is organized as follows. In Section 2, we discuss the related work. Section 3 details the process of building Indian Agricultural Research ontology. Section 4 brings out the relation of the developed ontology with Description Logics. Section 5 outlines the technical details of the developed ontology-based information retrieval system. While Section 6 describes various search options provided by the system, Section 7 brings out the importance of such systems for Digital Libraries. We conclude with a brief summary and our plans for future research in Section 7.

2 Related Work

Efforts are underway for developing ontology-based information retrieval systems for agriculture domain. The Food and Agriculture Organization of the United Nations (FAO) has made commendable contribution through the development of AGROVOC, a multilingual, structured and controlled vocabulary, which covers the terminology of all subject fields in agriculture, forestry, fisheries, food and related domains [4]. An attempt to convert the AGROVOC thesaurus into RDFS ontology was carried out during 2003 [5]. The FAO has also initiated the Agricultural Ontology Service (AOS), a reference tool that structures and standardizes agricultural terminology in multiple languages [6]. The AOS provides terms, definition and relationship components for sharing among associated partners and increasing the functionality for indexing and retrieving of resources. AGROVOC is being extensively used across the globe for developing multilingual agricultural thesaurus. Such thesaurus is being used for developing Semantic Web technologies based agricultural information systems. In the Indian scenario, the DEAL project [7] is currently in progress for developing ontology and a metadata system that supports the knowledge archiving & retrieval reuse in Indian Agriculture and rural livelihood domain. DEAL also proposes to develop a multilingual agricultural thesaurus based on AGROVOC. Angrosh and Urs [8] have successfully developed a prototype ontology-based information retrieval system for a specific case of Agri-Pest domain.

Most of the approaches for developing information systems employing AGROVOC focus on multilingual features and lay less emphasis on semantic relations. For instance Liang et al. [9] put forward a schema for mapping Chinese Agricultural Thesaurus to

FAO's AGROVOC. However, the crucial task of representing rich semantic relations of agricultural theses is absent. Though Sini et al. [10] have developed an ontology-based navigation system in the domain of food, nutrition and agriculture, the scope is limited to bibliographic metadata. Though such knowledge models facilitate semantic browsing, the real value of ontologies is obtained through use of rich semantic relations derived beyond bibliographic metadata relations. Thus, we present here a different methodology for developing ontologies based on the information available in the titles of the documents. We also show that the proposed methodology can be used for developing ontology-based information retrieval system.

3 Indian Agricultural Research Ontology

The field of Indian Agriculture is a cascade of many interacting effects between plant and environmental factors. Socio-economic factors such as timely-credit, price support, availability of critical inputs, crop insurance etc. play an important role in crop production. Post-harvest techniques such as storage and preservation mechanisms are also equally important. Thus the domain of Indian agriculture is a blend of various inter-disciplinary subjects and research in this domain interweaves these disciplines. Further, these research issues are specific to particular geographical regions. Thus, the Indian Agricultural Research ontology besides representing these interdisciplinary subject characteristics should also represent the geographical knowledge. Use of agricultural vocabularies such as AGROVOC would add value to the ontology. A broad framework of Indian Agricultural Research system mapping agricultural eTheses is shown in Figure 1.

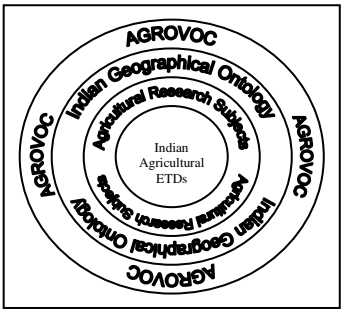

Fig. 1. Framework of Indian Agricultural Research Ontology

At a fundamental level, ontologies capture static domain knowledge in a generic way and provide a commonly agreed understanding of that domain, which may be reused and shared across applications and groups [11]. Thus, the primary task in developing ontology based information system for agricultural eTheses is to develop a shared understanding of the agricultural domain, which captures knowledge represented in agricultural ETDs. The data pertinent to agricultural ETDs in Vidyanidhi is of two types viz., full-text theses and bibliographic metadata. Vidyanidhi Digital Library and E-Scholarship Portal is set up at the University of Mysore for catering to the research

needs of the scholarly community in India [12]. Vidyanidhi currently hosts a full-text database of 6000 doctoral theses of various subjects and a bibliographic metadata database of more than 1 lakh eTheses records. There were nearly 200 full-text theses and 2800 agricultural eTheses records in the domain of agriculture at Vidyanidhi. The study focused on developing an ontology, using the limited semantic information available in the titles of agricultural records. The methodology followed for developing the ontology is as follows:

- Identify all agricultural eTheses titles in the Vidyanidhi Digital Library
- Identify all possible keywords in these titles. Keywords are primarily those terms that are used by a user for searching information.
- Identify and define classes and subclasses to which these terms belong to
- Define relations binding individuals of different classes.

Each of the 2800 agricultural eTheses titles was carefully analyzed to identify the possible keywords present in these titles. Upon identification of keywords, we identified the different classes and subclasses to which these keywords would belong. The classes and subclasses relationships were identified using agricultural handbooks and subject classification systems [13]. Table 1 shows a sample database of the identified keywords, classes and subclasses and relationships between keywords of different classes.

Table 1. Keywords, Classes, Subclasses and Relations identified in Agricultural eTheses

Sl. No.	Title	Class & Individuals	Class & Individuals	Class & Individuals	Geographical Class
1	Capital formation in arid agriculture: a study of resource conservation and reclamation measures applied to arid agriculture in Andhra Pradesh	Types of Agriculture (C) → Arid Agriculture (I)	Natural Resource Mgmt. (C) → Resource Utilization (I)	Agribusiness (c) → Agriculture Finance (c) → Capital Formation (I)	Indian States → Andhra Pradesh (I)
		Capital Formation inRelationTo Resource Utilization inRelationTo Arid Agriculture inRelationTo Andhra Pradesh			
2	Employment in Indian agriculture: Analytical and policy issue	Agricultural Economics → Agricultural Labour → Agricultural Employment (I)	Agricultural Policy → Agricultural Policy Issues		
		Agricultural Employment isRelated To Agricultural Policy			
3	Pattern of investment in agriculture in Orissa during the plan period	Agricultural Economics → Agricultural Investments → Pattern of Investments (I)	Agricultural Policy → Five Year Plans		Indian States → Orissa
		Agricultural Investments inRelationTo Five Year Plans inRelationTo Orissa			
4	Economic impact of central sector scheme women in agriculture on farm women in Maharashtra State	Agricultural Labour → Women Labour (I)	Agricultural Policy → Central Sector Schemes (I)		Indian States → Maharashtra
		Agricultural Women Labour inRelationTo Central Sector Schemes inRelationTo Maharashtra			

The agricultural research ontology was implemented in the OWL Web Ontology Language, a W3C recommendation for defining and instantiating Web Ontologies [14]. The Protégé-OWL editor was used for developing the ontology [15]. Figure 2 shows the screenshot of the ontology developed in Protégé-OWL Editor.

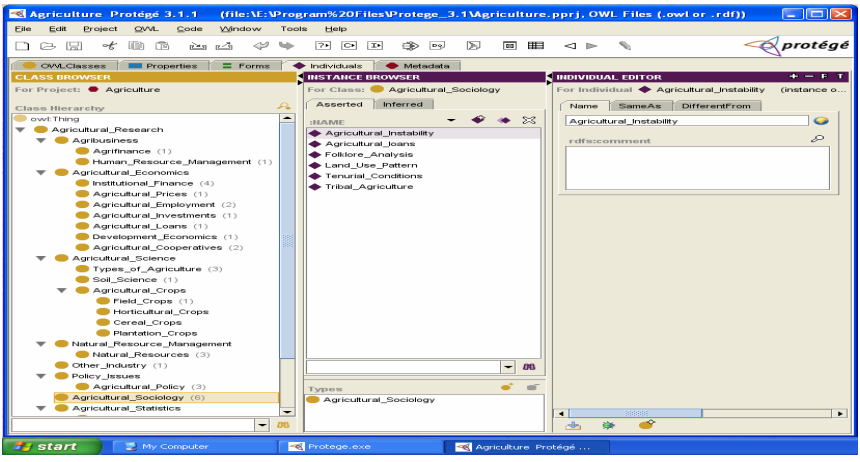

Fig. 2. Agricultural Research Ontology developed in Protégé-OWL Editor

A schematic representation of a part of Indian Agricultural Research ontology mapped with AGROVOC as is shown in Figure 3. These mappings provide rich valuable search points for retrieving agricultural ETDs.

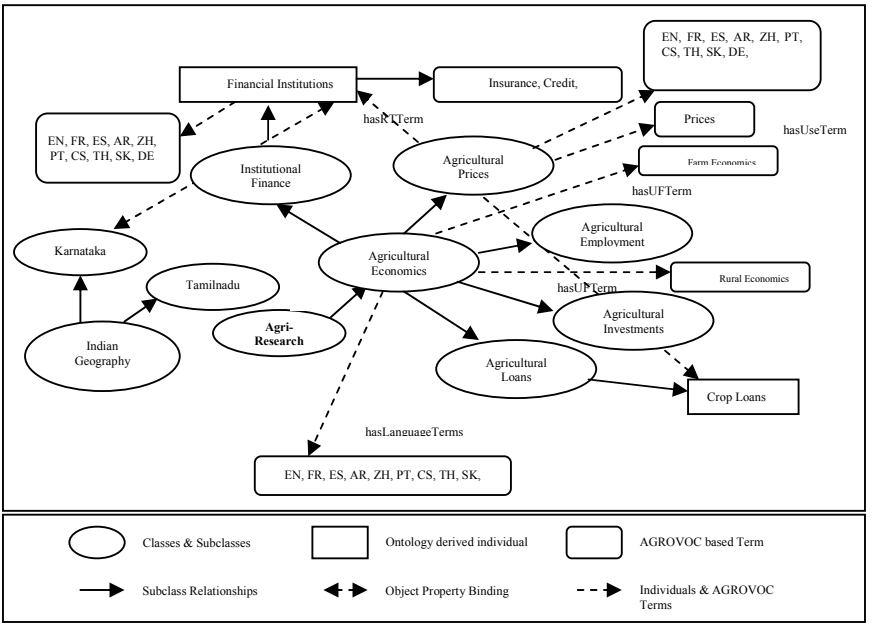

Fig. 3. Schematic Representation of Indian Agricultural Research Ontology

4 Description Logics Based Formalisms

The knowledge structures obtained above evolve into Description Logics (DLs) based knowledge representation (KR) formalisms. Primarily, DLs formalisms represent knowledge of an application domain by first defining the relevant concepts of the domain (its terminology) and then using these concepts to specify properties of objects and individuals occurring in the domain [16]. Description Logics based knowledge representation formalisms provide strong support for reasoning services, allowing inference of implicitly represented knowledge from the knowledge that is explicitly contained in the knowledge base. The subsumption relationships resulting in a hierarchical structure of classes and subclasses of the agricultural research domain can be used for designing value added information services. The classification and binding of individuals through relative object properties facilitate in deriving explicit knowledge about individuals associated with various classes.

The semantics of concept description in DLs is defined by the notion of interpretations, wherein an interpretation I consists of a non-empty set Δ^I (the domain of interpretation) and an interpretation function, which assigns to every atomic concept A a set $A^I \subseteq \Delta^I$ and to every atomic role R a binary relation $\subseteq \Delta^I \times \Delta^I$

$$\text{Interpretation } I = (\Delta^I, .^I)$$

DLs knowledge base typically comprises of two components viz. a "TBox" and an "ABox". While the TBox contains intensional knowledge in the form of a terminology, built through declarations that describe general properties of concepts, the ABox contains extensional knowledge – also referred to as assertional knowledge. Assertional knowledge refers to the knowledge that is specific to the individuals of the domain of discourse.

The Agricultural Research domain ontology developed in the study falls in line with DLs, with the keywords forming the assertional knowledge and the classes and subclasses forming the terminology of the domain. Further, the study, while conceptualizing the vocabulary of the knowledge base in terms of concepts and roles, maintained the important assumptions about DL terminologies, which included:

- allowance of only one definition for a concept.
- acyclic characteristics of the definitions – in the sense that concepts are neither defined in terms of themselves nor in terms of other concepts that indirectly refer to them

DLs based reasoners are employed for drawing inferences from the knowledge representation derived above. We used Pellet, a capable OWL-DL reasoner with acceptable to very good performance [17] for deriving inferences. The following section outlines the technical details of the developed system.

5 Technical Details of the System

The OWL-DL Agricultural Research ontology was used for developing ontology-based information retrieval system for Indian agricultural eTheses in Vidyanidhi Digital Library. The architecture of the system is as shown in Figure 4.

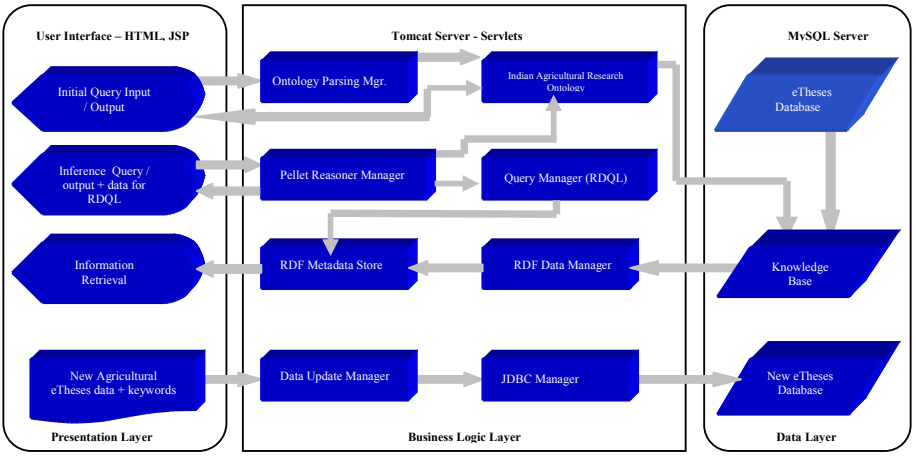

Fig. 4. Architecture of ontology-based IRS for Vidyanidhi Digital Library

5.1 Presentation Layer

The presentation layer mainly used HTML and JSP pages for creating user-friendly interfaces. The user interfaces are divided into two categories viz., Information Retrieval and Information Updating interfaces.

5.2 Business Logic Layer

The business logic layer has the following components:

5.2.1 Ontology Parsing Manager

The system uses Jena, an open source Java based API, developed by HP Labs for handling semantic web information model and languages [18]. The Jena2 ontology API is used to parse OWL for deriving class and subclass relationships and listing individuals.

5.2.2 Pellet Reasoner Manager

This module uses Pellet - a OWL-DL reasoner [17] to reason about individuals in the ontology. The reasoner is employed to retrieve related individuals connected by object-property relationships in the ontology. The retrieved individuals form the input for the Query Manager for retrieving respective agricultural eTheses.

5.2.3 Query Manager

The Query Manager is responsible for retrieving eTheses from an RDF data model, mapping agricultural eTheses and keywords defined in the ontology. The module uses Jena's Resource Description Query Language (RDQL) specific API function calls for querying the data model. The SQL-like syntax of RDQL is proved to be an effective way of querying an RDF data model [19].

5.2.4 RDF Data Manager

RDF Data Manager is primarily responsible for creating Resource Description Framework (RDF) Metadata store of agricultural eTheses, mapped with Indian Agricultural Research knowledge base. Jena's RDF API is used for representation of

models, resources, properties, literals, statements and other key concepts of RDF [20]. Table 2 shows RDF metadata capturing ontology defined keywords and Agrovoc vocabulary terms for a sample record.

Table 2. RDF metadata capturing keywords and Agrovoc vocabulary terms for a sample record

```
<rdf:RDF
xmlns:rss="http://purl.org/rss/1.0/"
xmlns:jms="http://jena.hpl.hp.com/2003/08/jms#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:owl="http://www.w3.org/2002/07/owl#"
xmlns:j.0="http://localhost:8080/vidyanidhi/india/agriresearch.owl#"
xmlns:ns3="http://localhost:8080/vidyanidhi/india/agriresearch.owl"
xmlns:daml="http://www.daml.org/2001/03/daml+oil#"
xmlns:ns2="http://purl.org/dc/elements/1.1/"
xml:base="http://localhost:8080/publication/documentowl" >
<rdf:Description rdf:about="http://localhost:8080/vidyanidhi/org/india/ETD1.htm">
    <j.0:hasKeyword1>Arid Agriculture</j.0:hasKeyword1>
    <ns2:title>Capital formation in arid agriculture: a study of resource conservation and reclamation
measures applied to arid agriculture</ns2:title>
    <j.0:hasKeyword2>Resource Utilization</j.0:hasKeyword2>
    <j.0:hasKeyword3>Capital Formation</j.0:hasKeyword3>
    <j.0:hasAgrovocBT1>Climate Zones</j.0:hasAgrovocBT1>
    <j.0:hasAgrovocNT1>Deserts</j.0:hasAgrovocNT1>
    <j.0:hasAgrovocRT1>Minimum Tillage</j.0:hasAgrovocRT1>
    <j.0:hasAgrovocRT2>Scrublands</j.0:hasAgrovocRT2>
    <j.0:hasTitle>Capital formation in arid agriculture: a study of resource conservation and reclamation
measures applied to arid agriculture</j.0:hasTitle>
    <j.0:hasAgrovocRT3>Dryland Management</j.0:hasAgrovocRT3>
    <ns2:creator>Jodha, Narpat Singh</ns2:creator>
    <ns2:contributor>Bardhan, Pranab</ns2:contributor>
    <ns2:contributor>Choudhury, Mrinal Dutta</ns2:contributor>
    <ns2:language>English</ns2:language>
    <ns2:degreeGrantor>University of Delhi</ns2:degreeGrantor>
    <ns2:year>University of Delhi</ns2:year>
    <j.0:hasAgrovocUF1>Drylands</j.0:hasAgrovocUF1>
    <j.0:hasAgrovocBT5>Resource Management</j.0:hasAgrovocBT5>
    <j.0:hasAgrovocNT5>Soil Conservation</j.0:hasAgrovocNT5>
    <j.0:hasAgrovocRT5>Natural resources</j.0:hasAgrovocRT5>
    <j.0:hasAgrovocRT6>Sustainable Development</j.0:hasAgrovocRT6>
    <ns2:description>The theses describes capital formation in arid agriculture</ns2:description>
    <rdf:type rdf:resource="http://localhost:8080/vidyanidhi/india/agriresearch.owl#hasIdentifier"/>
  </rdf:Description>
</rdf:RDF>
```

5.2.5 Data Update Manager

The Data Update Manager is responsible for adding new eTheses and ontology related keywords to the system.

5.3 Data Layer

The system employs MySQL database for creating the data backend, comprising of eTheses database and knowledge base created with the ontology.

6 Ontology-Based Information Search

The ontology-based information system facilitated the following search options:

6.1 Simple Search

The simple search option facilitates in performing search on terms present in the ontology, metadata of agricultural records and AGROVOC vocabulary. The generic search option retrieves all records related to a specific term by using the rich semantic relations binding metadata records with the ontology and the AGROVOC vocabulary.

6.2 Taxonomic View

A taxonomic view of the agricultural research domain is presented to the user. The user is provided with the option of browsing the subject hierarchy and view the instances (or keywords) of a specific class and retrieve records related to a specific keyword. This facility facilitated in confining user's search to a specific term.

6.3 Query Building Mechanisms

The system also facilitates in extending the taxonomic based information search and retrieval for query building mechanisms. The Pellet Reasoner is employed for developing such query mechanisms. For example, consider a specific individual, say 'Rice Crops' of the class 'Agricultural Crops' being searched. The system notifies the user that the individual 'Rice Crops' is related to Classes such as 'Agricultural Statistics', 'Agricultural Economics' etc. Further, if the user is interested in say, 'Agricultural Production', the system lists out various keywords such as 'Plant Development', 'Disease Control', etc., giving an option for the user to choose from the popup list. Furthermore, on specifying a specific keyword from the related class, the system notifies the available geographical entities related to the keyword. The user is again given the option for selecting the desired geographical location. Thus, this process results in building a chain of related keywords and retrieves records based on a specific chain, resulting in 'query building mechanisms'.

7 Implications of the Present Work

The process of ontology building primarily experiences difficulties in providing a significant coverage of the domain. It is also equally important to foster at the same time, the conciseness of the model by determining the meaningful and consistent generalizations [21]. The approach adopted in the present study overcomes these problems to a great extent. The development process of the ontology is relatively simpler and manageable. Further, while achieving the conciseness of the knowledge model, the identification of classes and subclasses provides a significant coverage of the domain. The ontology can be easily extended through addition of more keywords, classes and relationships. The ontology framework developed for Indian Agricultural Research domain is generic and is extendable for other domains in the Digital Library.

8 Conclusions and Future Work

In the present study, we have reported the development of an Indian Agricultural Research ontology based on the information available from the titles of agricultural eTheses. This was utilized for developing ontology-based information retrieval system for Vidyanidhi Digital Library. The work carried out in the present study is a novel effort for developing ontology-based information retrieval systems for digital libraries. The evolved framework and methodology can be extended for other domains as well. The use of AGROVOC increased the robustness of the system. Similar vocabularies in other domains can be used in the system. In future, we plan to target at extending the system for other domains in the Digital Library. This would result in the development of Indian Research ontology comprising various sub-domains. We also propose to investigate the use of vocabularies for different domains. Employing the same, we aim at developing a robust ontology-based information system covering all domains for Vidyanidhi Digital Library.

References

1. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley Longman Publishing (1999)
2. Guarino, N.: Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction and Integration. In: International Summer School - SCIE-97 on Information Extraction: A Multidisciplinary Approach to an Emeging Information Technology, pp. 139–170 (1997)
3. Ratsch, E., Schultz, J., Saric, J., Cimiano, P., Wittig, U., Reyle, U., Rojas, I.: Developing a Protein Interactions Ontology. Comparative and Functional Genomics 4(1), 85–89 (2003)
4. AGROVOC Thesaurus, FAO (2007), http://www.fao.org/aims/ag_intro.htm
5. Applied Ontologies in FAO.: FAO (2007), http://www.fao.org/aims/onto_domains.jsp
6. AGROVOC Concept Server.: FAO (2007), http://www.fao.org/aims/aos.jsp
7. Digital Ecosystem for Agriculture and Rural Livelihood Project.: (DEAL), Indian Institute of Technology Kanpur (2007), http://emandi.mla.iitk.ac.in/deal/
8. Angrosh, M.A., Urs, S.R.: Ontology-driven Knowledge Management Systems for Digital Libraries: Towards creating semantic metadata based information services. In: Proceedings of National Seminar on Knowledge Representation and Information Retrieval, Paper:N. Document Research & Training Centre, ISI, Bangalore (March 22-24, 2006)
9. Liang, A., Sini, M., Chun, C., Sijing, L., Wenlin, L., Chunpei, H., Keizer, J.: The Mapping Schema from Chinese Agricultural Thesaurus to AGROVOC, FAO (2000), ftp:// ftp.fao.org/docrep/fao/008/af241e/af241e00.pdf
10. Sini, M., Salokhe, G., Pardy, C., Albert, J., Keizer, J., Katz, S.: Ontology-based Navigation of Biliographic Metadata: Example from the Food, Nutrition and Agriculture Journal, FAO (2000), ftp://ftp.fao.org/docrep/fao/009/ah765e/ah765e00.pdf
11. Castano, S., Ferrara, A., Montanelli, S.: Dynamic Knowledge Discovery in Open, Distributed and Multi-Ontology Systems: Techniques and Applications. In: Web Semantics and Ontology, ch. 5, Idea Group Publishing (2005)
12. Vidyanidhi Digital Library and E-Scholarshop Portal.: University of Mysore (2007), www.vidyanidhi.org.in
13. Indian Council for Agricultural Research.: Handbook of Agriculture. Indian Council of Agricultural Research, New Delhi (2006)
14. Web Ontology Language (OWL), W3C (2004), http://www.w3.org/2004/OWL/
15. Protégé-OWL (ed.): Stanford Medical Informatics (2007), http://protege.stanford.edu/ overview/protege-owl.html
16. Baader, F., Knutt, W.: Basic Description Logics. In: Baader, F., Calvanese, D., McGuinness, D.L., Nardi, D., Patel-Schneider, P.F. (eds.) The Description Logic Handbook: Theory, implementation and applications, pp. 47–100. Cambridge University Press, Cambridge (2003)
17. Sirin, E., Parsia, B., Grau, B.C., Kalyanpur, A., Katz, Y.: Pellet: A practical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web 5, 51–53 (2007)
18. McBridge, B.: Jena: A semantic web toolkit. IEEE Internet Computing. November-December, 55–59 (2002)
19. Powers, S.: Practical RDF. O'Reilly (2003)
20. Min, W., Jianping, D., Yang, X., Xenxing, X.: The Research on the Jena-based Web Page Ontology Extraction and Processing. In: SKG 2005. Proceedings of the First International Conference on Semantics, Knowledge and Grid, IEEE Computer Society, Los Alamitos (2006)
21. Cimiano, P.: Ontology Learning and Population from Text: Algorithms, evaluation and applications. Springer, New York (2006)