

Ontology Development: A Case Study for Thai Rice

Areer Thunkijjanukij^{1*}, Asanee Kawtrakul²,
Supamard Panichsakpatana³ and Uamporn Veesommai⁴

ABSTRACT

This research reports on a pilot project that aimed to develop a prototype ontology for plant production using Thai rice as a case study. It is expected that the developed ontology will be used as a prototype model for other efforts to develop plant production ontology in the future. The Thai Rice Production Ontology provides an organizational framework of 2322 concepts and 5603 terms, in a system of hierarchical relations, together with 57 associative relations and 12 equivalence relations that allows reasoning about rice-production knowledge. The query expansion and reasoning components of the rice-production ontology can improve the performance of information retrieval and answer questions that a retrieval system without the ontology cannot. Terms in the ontology were used to query the Thai Rice Research Database (1350 records). The efficiency of the query was measured in terms of precision and recall. The experiment was conducted using five competency questions, and 93 queries were also defined. Retrieval experiments compared conventional search and ontological search approaches, supported with rice-production ontology-based query expansion. Results showed that the precision and recall rates increased on average from 0.08 to 0.72 and 0.01 to 0.64, respectively. The Thai Rice Production Ontology that has been developed will be a knowledge base for the management of knowledge on rice-production research in Thailand.

Key words: rice-production ontology, knowledge management, construction of ontology

INTRODUCTION

Research information is one of the critical factors for research development, both in terms of research policy formulation and enhancing researchers' capabilities. Therefore, all past studies and results of investigation are regarded as a valuable part of the knowledge base for research development. However, conventional search engines cannot interpret the sense of the

user's search, so not all the documents that discuss the search concept can be retrieved and often the ambiguity of the query leads to the retrieval of irrelevant information. Conventional search engines that match query terms against a keyword-based index will fail to match relevant information when the keywords used in the query are different from those used in the index, despite having the same meaning (Soergel *et al.*, 2004)

¹ Thai National AGRIS Centre, Kasetsart University, Bangkok 10900, Thailand.

² Department of Computer Engineering, Kasetsart University, Bangkok 10900, Thailand.

³ Department of Soil Science, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand.

⁴ Department of Horticulture, Faculty of Agriculture, Kasetsart University, Bangkok 10900, Thailand.

* Corresponding author, e-mail: libarn@ku.ac.th

A number of search engines are now emerging that use techniques to apply ontology-based domain-specific knowledge to the indexing including: similarity evaluation, results expansion and query enrichment processes. Ontology has been moving from the domain of artificial-intelligence laboratories to the desktops of domain experts. Many ontologies have been developed, such as Rice Ontology (RO), *Zea mays* Ontology and some ontologies in the NeOn Project. Rice ontology is an ontology that specializes in the genome informatics of rice and has been developed as a biological-domain ontology for exchanging genome informatics (Takeya *et al.*, 2003). *Zea mays* ontology presents plant structure, included the anatomy and morphology of maize and also comprises international botanical terms, references, synonyms and phylogenetic information (Vincent *et al.*, 2003). NeOn is a European project that aims to advance the state of the art in using ontologies for *large-scale* semantic applications in distributed organizations and improve the capability to handle multiple-networked ontologies that exist in a particular context (<http://www.neon-project.org/web-content/>).

An important role of ontologies is to serve as schemata or 'intelligent' views over information resources. Thus, they can be used for indexing, querying and reference purposes over non-ontological datasets and systems (Davies *et al.*, 2006). In the context of computer and information sciences, an ontology defines a set of representational primitives with which to model a domain of knowledge or discourse. The representational primitives are typically classes (or sets), attributes (or properties) and relationships (or relations among class members). The definitions of the representational primitives include information about their meaning and constraints on their logically consistent application (Gruber, 2007).

The construction of an ontology on a

specific plant or crop has not yet been reported in the literature. Consequently, the objectives of this pioneer and pilot-work research were to: develop an ontology prototype for plant production using rice production as a test case study; and apply the ontology to an information retrieval mechanism, as a knowledge base for retrieving and managing knowledge, in the domain of agriculture. This ontology prototype will be a model for the future development of other agricultural ontologies, so that management of agricultural knowledge will be more efficient.

MATERIALS AND METHODS

Materials

The knowledge resources used for developing the rice production ontology consisted of: 67 rice production and related-subject textbooks; 17 related websites; the Thai AGROVOC Thesaurus (Thai National AGRIS Centre, 2004); and the Thai Rice Research Database, with 1350 records indexed using the Thai AGROVOC Thesaurus (Thai National AGRIS Centre, 2008). In addition, ontology applications for construction and visualization were used, such as the FAO AGROVOC Concept Server Workbench tool (<http://www.fao.org/aims/agrovocs.jsp>) for ontology construction, Touch Graph for visualizing the developed ontology, CmapTools version 4.08 COE (<http://cmap.ihmc.us/>) and MindManager ver. X5 for knowledge modeling.

The research task for the construction of the rice production ontology was divided into five stages: ontology specification; knowledge acquisition; conceptualization; formalization and implementation.

Methods

Ontology specification

The ontology domain and scope were designed by sketching two kinds of questions:

“basic questions” (Noy and McGuinness, 2001) and “competency questions” (Gruninger and Fox, 1995). The basic questions, which clarified the purpose of the ontology and limited the scope of the model, consisted of examples such as: What is the domain that the ontology will cover? What are we going to use the ontology for? What types of questions will the information in the ontology provide answers for? Who will use and maintain the ontology?

Competency questions are lists of questions that a knowledge base associated with the ontology should be able to answer. These competency questions should just be a sample and do not need to be exhaustive. The answers to these questions may change during the ontological design process. Five competency questions were collected by interviewing the managers of rice research projects in the Rice Department. These questions were used to create queries for evaluation by comparing the results from a keyword-base search (conventional search) and an ontology-based query expansion (ontology search). The questions are listed below:

a) Jasmine rice is the most popular rice variety of Thailand. How many jasmine rice research references in the literature are defined by each subject from well-known classification schemes?

Relations to use for query expansion are: hasSynonym, hasTranslation.

b) How many research references focus on rice biological control organisms?

Relations to use for query expansion are: hasBiologicalControlAgent, hasCommonName, hasSynonym, hasTranslation.

c) What is the most common rice disease research in Thailand?

Relations to use for query expansion are: hasDiseases, hasPathogen, , hasSynonym, hasTranslation.

d) How many rice research papers contain chemical fertilizer and organic fertilizer?

Relations to use for query expansion are: hasSubClass, hasSynonym, hasTranslation.

e) How many research papers are concerned with rice pest control, divided by type of pest, namely “field pest” and “stored product pest”?

Relations to use for query expansion are: hasPest, hasRelatedType [field pest], hasRelatedType [stored product pest], hasSubClass, hasCommonName, hasSynonym, hasTranslation.

The terms of the concepts used to create queries are preferred terms and all of the synonyms that represent that concept (both in Thai and English) such as: acronym, abbreviation term, spelling variance term, singular/plural, chemical symbol, trade name, common name, local name, etc.

Knowledge acquisition

The methodology used for this approach was a combination of text analysis and an expert approach. The first step was to extract as much plant production-based knowledge as possible from the literature and to collect and review the related knowledge resources and categorize them systematically. The categories should cover all topics related to rice production from the starting process of cultivation to harvesting, including rice pest protection and rice breeding. Since rice production is related to many disciplines, there was a need to collect knowledge comprehensively from multiple resources. Domain-specific knowledge was captured from both the explicit knowledge (knowledge that can be written down, shared with others and stored in a database, such as: reports, procedures, instructions), and tacit knowledge (knowledge that resides inside people, such as: experiences, intuition, insights) of the experts. Interviews and discussion were some of the techniques used to acquire knowledge from experts.

The second step was to analyze and summarize the knowledge. This step involved the

study of all of the knowledge sources from the first step, which had been summarized and organized in a structural form. A final revision by experts confirmed the data structure.

Conceptualization

A conceptual ontological model consists of the concepts in the domain and the relationships among those concepts. Conceptual modeling involves defining the ontological model structure, identifying concepts, identifying relationships and then creating informal draft models using the previous summarized knowledge and knowledge modeling tools, such as MindManager and CmapTool.

The rice production ontology collected and combined rice with multiparts, effecting factors by many categories of relationship. The ontology was formed by analyzing plant production knowledge, based on the “Whole Plant Model” by Beverly *et al.* (1993).

All relationship names are written starting with lower case and capitalizing other words, without any spaces (Sini and Yadav, 2009). For hierarchical relationships, there is only one relation namely “hasSubclass” (Figure 1). This relation is defined between all of the hierarchical concepts. Associative relationships are defined by identifying relating verbs between concepts and assigning a relation name to form a meaningful statement. The most common way to label a relationship is by role names.

Formalization

The rice ontology conceptual model from the previous step was transformed into a formal model by writing it in a formal form. The steps to convert a conceptual model to a formal form are: listing all the concepts and relationships in the conceptual model in a data sheet; defining terms that represent concepts by selecting a preferred term as a concept representative, so that non-preferred terms will be assigned as a synonym; defining terminology relationships; and defining concept properties.

All relationship names should be written starting with lower case and capitalizing other words, without any spaces (Sini and Yadav, 2009). There are three types of terminology relationships.

1) Concept-to-term relationship, namely “hasLexicalization”. This is the relationship between the concept and the selected preferred term. For example: concept[rice] hasLexicalization term[*Oryza sativa*].

2) Term-to-term relationships. All of these relations are used for preferred terms and their synonyms, which are difference terms, such as: hasAcronym, hasAbbreviation, hasSpelling Variant, hasPural or hasSingular, hasCommon Name, hasLocalName, hasScientificName, hasTradeName, hasChemicalSymbol, has ChemicalFormula, hasTranslation, hasSynonym. For example: term[rice] hasPural term[rices], term[sulphur] hasSpelling Variant term[sulfer],

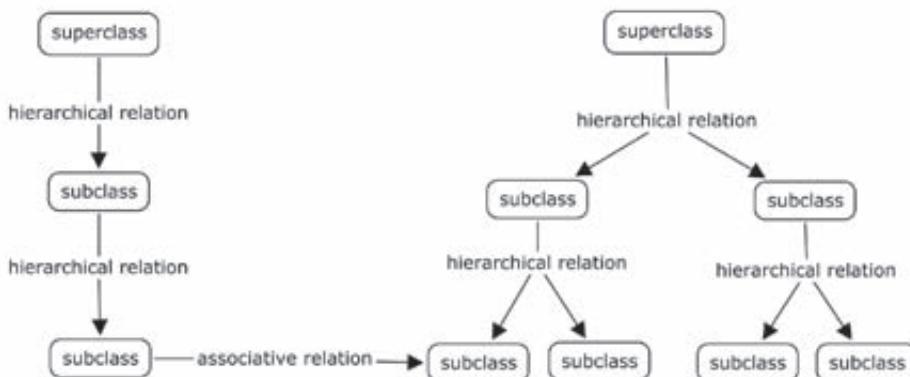


Figure 1 Ontology structure model.

term[*Oryza sativa*] hasCommonName term[rice].

3) Concept-to-concept relationship.

These relations connect concepts (represented by the preferred term) in a different position in the hierarchy, such as: hasPest, hasDisease, hasPathogen, hasRelatedType, etc

Implementation

This research implemented a formalized rice production ontology using the FAO AGROVOC Concept Server Workbench Tool (AGROVOC CS WB) (FAO, 2008; <http://www.fao.org/aims/agrovoccs.jsp>) for knowledge representation in the form of OWL DL (Liang *et al.*, 2006). Concepts and relations were formalized and verified as a datasheet table (XLS format). A converting application was developed by the Naist Laboratory, Department of Computer Engineering, Kasetsart University, to transfer data from the Excel datasheet to the AGROVOC CS WB format.

Ontology visualization in AGROVOC CS WB was implemented as an added-on module, which was developed for the Thai rice production ontology. The “Thai Agricultural Ontology Visualization Tool” was developed from the open source “Touch Graph” (<http://sourceforge.net/projects/touchgraph>) by the Thai National AGRIS Centre, Kasetsart University in consultation with the Naist Laboratory. The tool was modified to visualize concepts and relations using both the Thai and English languages.

Ontology evaluation process

The quality of the rice production ontology was judged by two methods. The first was validation by experts and the second was evaluation by users. The domain-specific experts verified the ontology and corrected it if needed. This step provided evaluation in terms of the theoretical correctness of the concepts, terms and relationships relevant to rice production.

The evaluation by users judged how good the ontology was in satisfying the competency questions, defined in the previous specification process by the research project managers. The

terms in the ontology were used to query the Thai Rice Research Database. The efficiency of the use of the ontology was measured in terms of the precision and recall of query search results by the three domain-specific experts.

Precision is a measure of exactness or fidelity, whereas recall is a measure of completeness. In an information-retrieval scenario, precision is defined as the number of relevant documents retrieved by a search, divided by the total number of documents retrieved by that search (Equation 1), and recall is defined as the number of relevant documents retrieved by a search, divided by the total number of existing relevant documents (which should have been retrieved) (Equation 2).

$$\text{Precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}|}{|\{\text{documents retrieved}\}|} \quad (1)$$

$$\text{Recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{documents retrieved}\}|}{|\{\text{relevant documents}\}|} \quad (2)$$

RESULTS

The Rice Production Ontology (RPO) was constructed from scratch in consultation with domain experts. The RPO covers the domain of rice production from cultivation to harvesting. Relevant knowledge related to rice production was analyzed, in particular using: 65 text books and 17 websites; the Thai AGROVOC Thesaurus; and consultation with 27 experts in this field. Concepts and relations were formalized and verified in a datasheet and imported into the AGROVOC Concept Server Workbench tool. A Thai Agricultural Ontology Visualization tool and an Ontology Tree Editor were developed to present the ontology as a graph in order to facilitate editing. Refinement in the loop involved performing the transformation with the criteria validated by an expert to improve the ontology created. The rice production ontology contains 2322 concepts and

5603 terms in hierarchical system with 57 different types of associative relations and provides an organizational framework that allows reasoning about rice production knowledge. Guidelines and criteria, together with rules for maintaining the ontology were created but they are not presented in this paper.

Ontology visualization

Visualization tools, such as Prefuse and Touch Graph, which are open-source tools, were selected to display this ontology. Prefuse was adjusted to display the ontology as an overview. Touch Graph was used to develop the Thai Agricultural Ontology Visualization Tool and connected to the AGROVOC Concept Server Workbench to present the Thai Rice Production

Ontology. The graph can be visualized in Thai or English; it is necessary to select the target concept and then click the visualized function. Moreover, the display can use either a hierarchical or a vertical view, and the user can retrieve information from a predefined database by a search that uses the concept selected in the graphical view (Figures 2 and 3). All the terms of that concept or the whole subclasses will be generated from the ontology and sent to the search mechanism in the connected database.

The Thai Agricultural Ontology Visualization Tool is connected to a search function. By right clicking on the target concept, the user can select the function “Search” (Figure 4). Two options have been implemented: search by the selected concept; or search by the selected

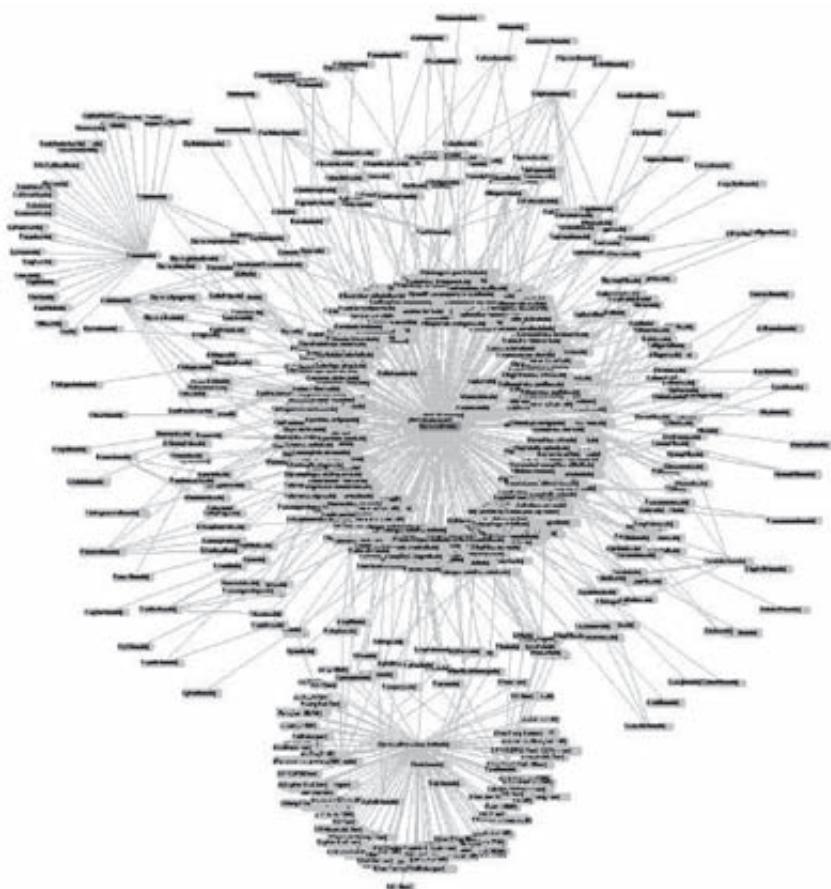


Figure 2 Thai Rice Production Ontology in the full view display.

concept and all subclasses. The terms of the selected concept will be generated and sent to query the Thai Rice Research Database; the results are displayed with metadata.

Rice production ontology evaluation results

The rice production ontology was evaluated with regard to its capabilities to satisfactorily handle the competency questions. Terms in the ontology were used to query the Thai Rice Research Database (containing 1350 metadata records). The retrieval efficiency was measured in terms of its precision and recall.

The experiment was conducted using five competency questions; 93 queries were defined. The retrieval experiment compared a keyword-based search (conventional search) and name-entity representation supported with ontology-based query expansion (ontology search). The results showed that the precision and recall rates increased on average from 0.08 to 0.72 and 0.01 to 0.64 respectively (Table 1).

DISCUSSION

Knowledge management systems or domain-independent applications use knowledge bases built in a form of ontology. An ontology for a specific domain is not a goal in itself. Developing an ontology has as its objective defining a set of data, which specific programs may use. The purpose of this research was to develop a rice-production ontology to be used as background knowledge for agricultural research-knowledge management systems. The quality of the ontology only can be assessed by using it in applications for which it was designed.

The Rice Production Ontology provided 2322 concepts and 5603 terms in a hierarchical system with 57 different types of associative relations. More than half of the concepts were object-entity concepts consisting of plant, pest animals, diseases organisms and agricultural substances, such as pesticides and fertilizers. A minor number were functional entity concepts,

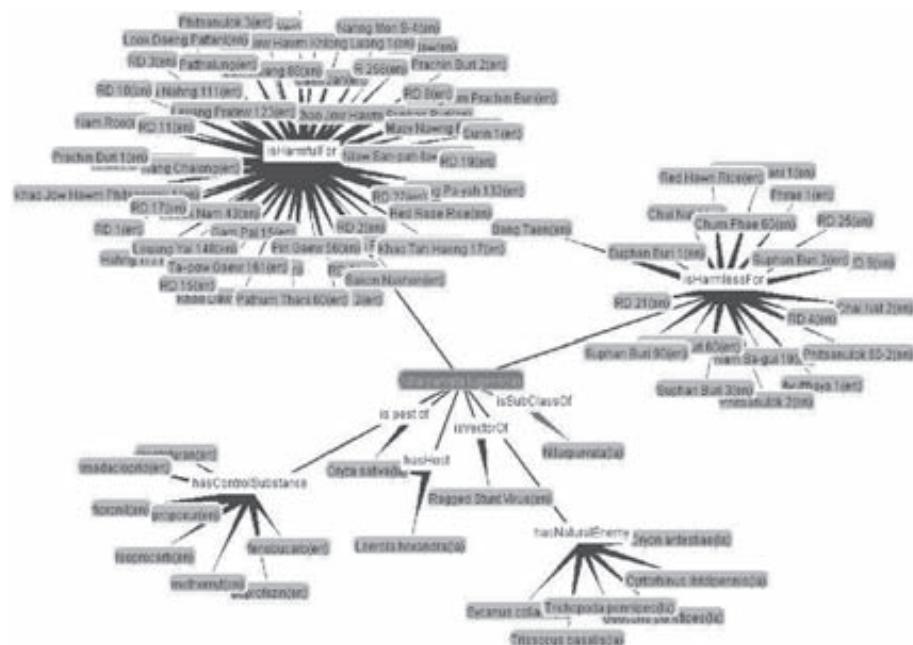


Figure 3 Thai Rice Production Ontology concept of the brown planthopper, showing the hierarchical (isSubclassOf) and associative (isPestOf, hasHost, isVectorOf, hasControlSubstance, hasNaturalEnemy, isHarmfulFor, isHarmlessFor) relationships.

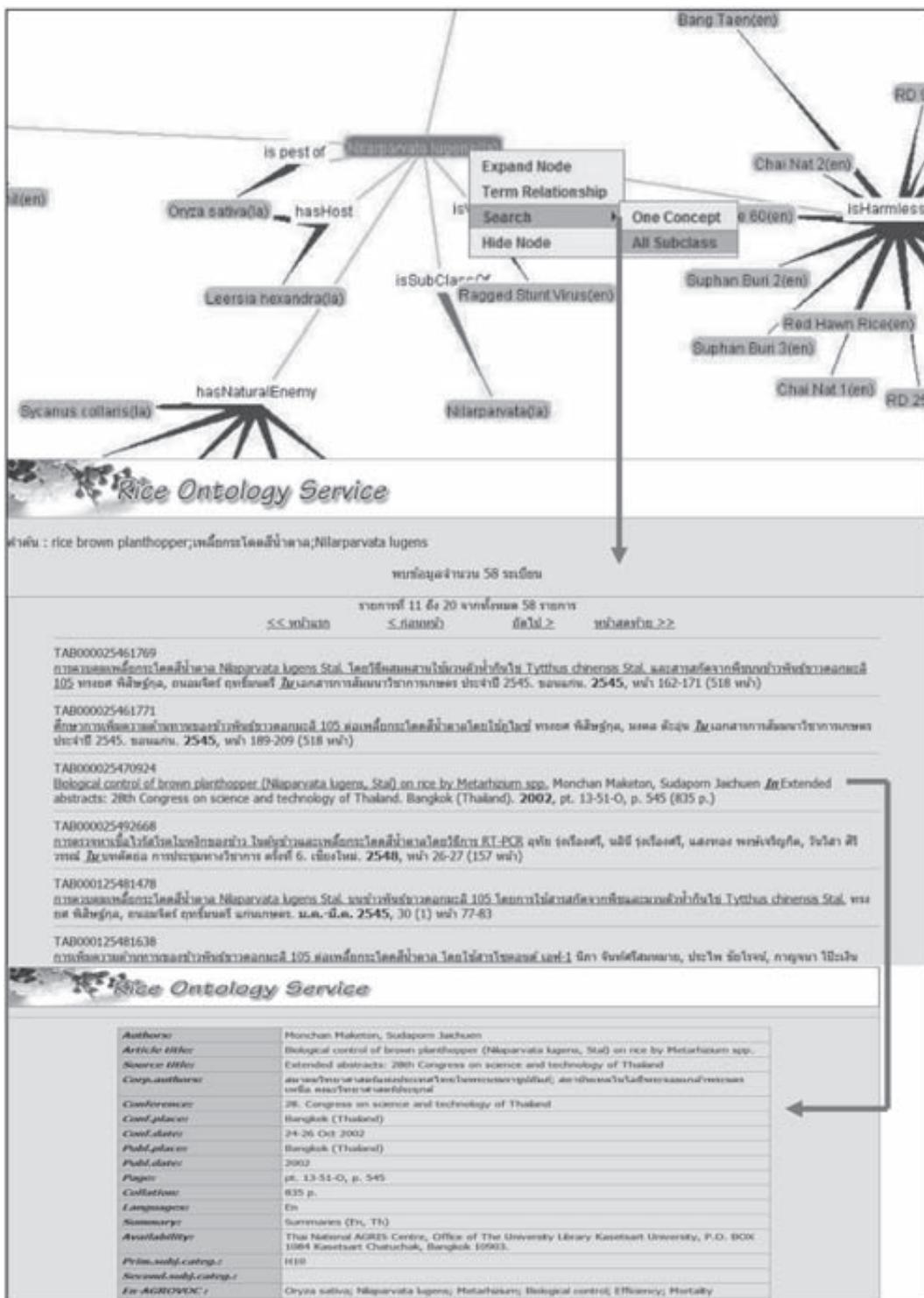


Figure 4 Thai Rice Production Ontology Visualization with search function and search results.

Table 1 Query results compared between a conventional search approach and an ontology-based search approach.

Approach	Search result (Average)				
	Relevant	Retrieved	Retrieved and Relevant	Precision	Recall
Conventional search	22.95	1.98	1.69	0.08	0.01
Ontology search	22.95	21.15	18.41	0.72	0.64

which indicated the function of rice production processes. The rice-production ontology was created as a skeleton ontology to be used in knowledge management systems associated with rice-production research; therefore, the class hierarchy was not too detailed. The ontology was designed to cover all topic-related concepts comprehensively and not to need additional specialization beyond that required by the application. For example, the ontology includes content related to rice morphology. This category contains few concepts concerning the vegetative and reproductive part of rice, which is related to production. This ontology does not include all the properties of rice but presents the most relevant properties, such as rice cultivar and resistance to pest or diseases. The rice cultivar characteristics (described with other information such as size, color, growth rate, etc.) have been omitted. In addition, the associative relationships between concepts have not been added for whole concepts; relations between the reactions of pesticides with each pest have not been included. The interconnections between concepts are defined for the scope and purpose of the ontological use only.

Retrieval efficiency described by the precision and recall rate was used to evaluate the ontology. In an information retrieval context, precision and recall were measured in terms of a set of retrieved documents and a set of relevant documents. Precision is a measure of exactness or fidelity, whereas recall is a measure of completeness. A perfect precision score of 1.0 indicates that every result retrieved by a search was relevant, whereas a perfect recall score of 1.0 indicates that all relevant documents were

retrieved by the search. The rice production ontology is capable of exploring the query topics and increasing the search results to greater extent than a conventional search can. This can be confirmed by the precision and recall scores of the ontological search, which were significantly higher than for the conventional search. From the search results, the number of relevant records from some queries was greater than for retrieved records. This was because concepts and terms collected in the ontology were not comprehensive enough. At the same time, the number of relevant queried records was less than for retrieved records. This meant that the system had queried irrelevant results. From the results, most of the irrelevant results were retrieved from abstracts, while the relevant results were queried from titles and keywords. To reduce the irrelevant results the search could be improved by searching only in titles and using well-defined keywords, or by applying more semantic techniques in the retrieval system. In addition, to provide more comprehensive search results to serve the users' needs, enrichment of the Thai Rice Research Database could be considered as one of the important issues.

CONCLUSION

The Thai Rice Production Ontology provides an organizational framework with 2322 concepts and 5603 terms in a hierarchical system, with 57 associative relations and 12 equivalent relations, that allows reasoning about rice-production knowledge. The relationships in the rice-production ontology were compared with the

existing relationships in the AGROVOC CS; 19 relationships were the same as in the AGROVOC CS and 51 new relationships were defined. Having compared all the concepts from the rice-production ontology with the existing terms in the FAO AGROVOC Thesaurus, it was concluded that 2687 terms (about 48%) in the ontology already existed in the Thai AGROVOC Thesaurus.

Concepts and relations were formalized and verified in the form of a datasheet and imported into the AGROVOC Concept Server Workbench tool. A Thai Agricultural Ontology Visualization tool and an Ontology Tree Editor were developed to present the ontology graphically and to assist ontology editors in their tasks. Refinement in the loop involved performing the transformation with the criteria validated by an expert to improve the ontology created..

The evaluation of the rice-production ontology involved identifying how extensively the ontology could be used to answer the competency questions. The query expansion for rice-production ontology could increase information retrieval efficiency and answer questions, which a traditional retrieval system without ontology could not do. Terms in the ontology were used to query the Thai Rice Research Database (1350 records). The efficiency of the query was measured in terms of its precision and recall rate, with the experiment conducted using five competency questions, in which 93 queries were defined. The retrieval experiments compared a conventional search and an ontology-based search supported with the rice production ontology-based query expansion. The results showed that precision and recall increased on average from 0.08 to 0.72 (nine times) and 0.01 to 0.64 (64 times) respectively, which indicated that the ontology-based search was more efficient than the conventional search.

This research should support knowledge service organizations, research planning sections and research project managers in making decisions using a knowledge base and in creating research-

knowledge management initiatives. This research effort also helped to establish ontology construction, increasing the efficiency of research information retrieval systems and enhance service quality for research-knowledge management efforts.

Finally, this research demonstrated that ontology plays a critical role in knowledge acquisition and knowledge management processes. It helps make knowledge storage and retrieval process significantly more intelligent. Thus, it is very reasonable to encourage the construction of many more ontologies. What then follows is the need for tools that improve the efficiency of constructing new ontologies, by transferring and merging existing ones.

Developing domain-specific ontologies is the biggest challenge for good information retrieval and knowledge services. Therefore, it is advisable that experts and information specialists in each specific knowledge domain should collaboratively start developing their respective ontologies. Furthermore, collaboration and cooperation among related organizations or ontology editors should be established, so that the developed ontology could be reused and be interoperable for substantial development.

ACKNOWLEDGEMENTS

The authors are grateful for all the useful comments from domain specific experts, especially the professors, agricultural subject experts in the Faculty of Agriculture, Kasetsart University, and information and knowledge management officers in the Knowledge and Communication Department, Food and Agriculture Organization. The authors would like to thank the programmers in the Department of Computer Engineering who gave technical support for data conversion. The authors are exceptionally thankful to the Thai National AGRIS staff, who provided support in data collection and

encouragement. The anonymous reviewers are also thanked for their careful revision and most valuable suggestions.

LITERATURE CITED

Beverly, R. B., J. G. Latimer, and D. A. Smittle. 1993. Preharvest physiological and cultural effects on postharvest quality, pp. 74-98. In R. L. Shewlelt and S. E. Prussia, (eds.). **Postharvest Handling: A Systems Approach.** Academic Press, Inc., San Diego.

Davies, J., R. Studer and P. Warren. 2006. **Semantic Web Technologies: Trends and Research in Ontology-based Systems.** John Wiley & Sons, England.

FAO. 2008. **AGROVOC Concept Server.** Agricultural Information Management Standards. Available Source: <http://www.fao.org/aims/agrovocs.jsp>, June 25, 2008.

Gruber, T. 2007. **Ontology.** Ontology (Computer Science) definition in Encyclopedia of Database Systems, Springer-Verlag. Available Source: <http://tomgruber.org/writing/ontology-definition-2007.htm>, January 15, 2008.

Gruninger, M. and M.S. Fox. 1995. Methodology for the design and evaluation of ontologies, pp 51-60. In **Proceedings of the Workshop on Basic Ontological Issues in Knowledge Sharing.** IJCAI-95, Montreal.

Liang, A.C., B. Lauser, M. Sini, J. Keizer and S. Katz. 2006. **From AGROVOC to the Agricultural Ontology Service/Concept Server An OWL model for managing ontologies in the agricultural domain.** Available Source: http://owl-workshop.man.ac.uk/acceptedPosition/submission_31.pdf

Noy, N. F. and D. L. McGuinness. 2001. **Ontology Development 101: A Guide to Creating Your First Ontology.** Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Standford University.

Sini, M. and V. Yadav. 2009. Building **Knowledge Models for Agropedia Indica Requirements.** Guidelines v. 1.0. Available Source: http://agropedia.iitk.ac.in/km_guidelines.pdf, January 12, 2009.

Soergel, D., B. Lauser, A. Liang, F. Fisseha, J. Keizer and S. Katz. 2004. Reengineering thesauri for new applications: the AGROVOC example. **Journal of Digital Information** 4(4). Article No. 257, 2004-03-17. Available Sources : <http://jodi.tamu.edu/Articles/v04/i04/Soergel/> November 15, 2007.

Takeya, M., H. Numa and K. Doi. 2003. Ontology Using Role Concept Recognized on Biological Relationships and Its Application. **Genome Informatics** 14: 685-686.

Thai National AGRIS Centre. 2004. **Thai Agriculture Thesaurus.** Thai AGROVOC. Available Source: <http://pikul.lib.ku.ac.th>, January 12, 2008.

Thai National AGRIS Centre. 2008. **Thai Rice Research Database.** Available Source: <http://pikul.lib.ku.ac.th/rice1>, January 12, 2008.

Vincent, P.L.D., E.H. Coe and M.L. Polacco. 2003. *Zea mays* ontology—a database of international terms. **Trends in Plant Science** 8 (11): 517-520.