Proc. Int. Conf. on Dublin Core and Metadata for e-Communities 2002: 113-123

A Comprehensive Framework for Building Multilingual Domain Ontologies: Creating a Prototype Biosecurity Ontology

Boris Lauser, Tanja Wildemann, Allison Poulos,
Frehiwot Fisseha, Johannes Keizer, Stephen Katz
Food and Agriculture Organization of the UN, GILW
Rome, Italy
{boris.lauser, tanja.wildemann, allison.poulos,
frehiwot.fisseha, johannes.keizer, stephen.katz}@fao.org
http://www.fao.org

Abstract

This paper presents our ongoing work in establishing a multilingual domain ontology for a biosecurity portal. As a prototypical approach, this project is embedded into the bigger context of the Agricultural Ontology Service (AOS) project of the Food and Agriculture Organization (FAO) of the UN. The AOS will act as a reference tool for ontology creation assistance and herewith enable the transfer of the agricultural domain towards the Semantic Web. The paper focuses on introducing a comprehensive, reusable framework for the process of semi-automatically supported ontology evolvement, which aims to be used in follow-up projects and can eventually be applied to any other domain. Within the multinational context of the FAO, multilingual aspects play a crucial role and therefore an extendable layered ontology modelling approach will be described within the framework. The paper will present the project milestones achieved so far: the creation of a core ontology, the semiautomatic extension of this ontology using a heuristic toolset, and the representation of the resulting ontology in a multilingual web portal. The reader will be provided with a practical example for the creation of a specific domain ontology, which can be applied to any possible domain. Future projects, including automatic text classification, and ontology facilitated search opportunities, will be addressed at the end of the paper.

Keywords: *Ontology, Semantic Web, Ontology creation, Ontology Engineering Framework, Ontology Learning, Multilingual Ontology, Biosecurity, Food Safety, Animal Health, Plant Health.*

1. Introduction

1.1 Motivation and subject domain

The management of large amounts of information and knowledge is of ever increasing importance in today's large organizations. With the ongoing ease of supplying information online, especially in corporate intranets and knowledge bases, finding the right information becomes an increasingly difficult task. Today's search tools perform rather poorly in the sense that information access is mostly based on keyword searching or even mere browsing of topic areas. This unfocused approach often leads to undesired results. The following example illustrates the problem more clearly:

One might, for example, want to find out which organization established the Agreement of Agriculture. A simple search for "establish Agreement of Agriculture" might result in a huge list of documents containing these words, but actually none of them containing the desired result: WTO or World Trade Organization. The problem becomes even worse, if the result searched for only appears in a foreign language document. Figure 1 shows an extract of an ontology, which could solve this problem. The grey ellipses represent generic concepts, whereas the white ones represent specific instances of these concepts. The two concepts shown here are interlinked by a relationship. The ontology enabled search application would first identify "Agreement of Agriculture" as a "standard" and would then detect the relationship "establish" to "international organization"

and its instances, and hence solve the problem by extending the search query. Furthermore, it could provide added value by detecting other relationships that provide the user with more possibilities, for example standards of other organizations could be presented.

This example shows how ontologies can help to improve the management of information. Semantically annotated documents, i.e. documents which are indexed with ontological terms and concepts instead of simple keywords, provide several advantages. First, the ontological abstraction provides robustness against changes in the document. In the above example, the document content might change using the term 'Agricultural Agreement' instead of 'Agreement of Agriculture'. However, since the document has been annotated with the ontological semantics, this will not affect the search results. Second, since the ontology used for annotating the document is domain specific, the semantic meanings and interpretations of keywords are bound to that domain and therefore the retrieval is likely to be more efficient. A term can have several meanings in different domains. By first mapping the keyword to its semantic representation in a specific ontology and using the ontology's linked knowledge structure, a much more focused search approach can be taken. Third, document specific representations no longer affect the search. This is extremely important in the case of multilingual representations. Keywords of several languages are mapped to the same concept in an ontology and are therefore given the same meaning. Multilingual search portals can be established to produce the same results, no matter which language is used for retrieval.

Another important issue of knowledge management, especially with regards to document metadata and indexing, is the classification of documents. Presently, this is carried out by subject specialists in a time consuming process. With today's vast amount of available information on the WWW, automatic support is needed to efficiently manage this task. Ontologies play a critical role in supporting the machine readable semantics needed to facilitate automation.

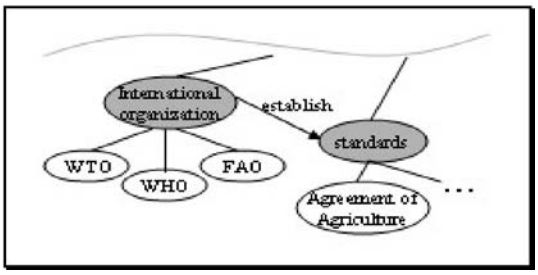

Figure 1. Ontology example, excerpt

Before such powerful Semantic Web[1] applications can be built and used within certain domains of knowledge, the basic requirement, a machine readable vocabulary represented by a domain ontology has to be established. The creation of ontologies is a time consuming task and often carried out in an ad-hoc manner. Only few methodologies exist, and even less automated tool support is available. Constituting the knowledge base for future Semantic Web applications, domain ontologies have to be created continuously in all possible areas and communities. The need for a reusable methodology is evident. This paper outlines a comprehensive, reusable framework for semi-automatically-aided building of domain ontologies. A prototype project is used for the application of this computer-aided framework, which provides the reader with a practical, methodological ontology engineering approach.

The domain that serves for creating the prototype ontology is the Biosecurity Portal on food safety, animal and plant health. The portal is an access point for official national and international information relating to biosecurity, the risks associated with agriculture (including fisheries and forestry), and food production. Many countries are still struggling with rapid advances in technology and often lack access to basic information on food safety, animal health and plant health. However, access to this information is of paramount importance for countries to protect health, agriculture and the environment.

One of the goals of the portal is to serve as an electronic information exchange mechanism for the addressed community and therefore to ensure efficient and effective information retrieval. The extension of its knowledge base to information available on various other sources in the WWW can highly support the purpose of the portal. Serving an international community, the information must be retrievable in various languages. The domain is multidisciplinary across three different, but related subject areas. The motivation to create a commonly agreed on, formally specified vocabulary in form of domain ontologies becomes evident

1.2 Overview of the approach

The presented project introduces a comprehensive framework for building a domain-specific ontology. The approach combines classical methodologies for human-based ontology engineering with semiautomatic support of a heuristic toolset. Actually, two methods for ontology acquisition are applied in order to create the domain ontology. The first is to create a small domain-specific core ontology from scratch and then apply a focused web crawler to this ontology in order to retrieve domain related web pages and interesting domain terms for extending this base

[1] Refer to (Palmer 2001) for an introduction to the Semantic Web.

ontology. The second acquisition approach takes a well-established thesaurus as a basic vocabulary reference set and converts it to an ontology representation. Then a domain specific and a general corpus of texts are used in order to remove concepts that are not descriptive for the domain. The heuristic used here is, that domain specific concepts are more frequent in the domain-specific text corpus. A side product of this removal step is again a list of frequent terms, which can eventually enhance the ontology (see Volz 2000 for more details on this approach). The results of these steps are assessed to assemble the final domain specific ontology, which is now accessible through a multilingual web portal.

1.3 Outline

The next section provides a brief introduction to the larger framework the prototype project is embedded in. In Section 3 a proposed layered multilingual ontology model is introduced. It sets the basis for the methodological framework, which is discussed in detail in Section 4. All steps of the prototype project are then presented in Section 5 and currently available results are shown. Finally, Section 6 gives an outlook on further work and opportunities that this project enables.

2. The project framework: FAO and the AOS

The Food and Agriculture Organization (FAO) of the United Nations (UN) is committed to helping combat and eradicate world hunger. Information dissemination is an important and necessary tool in furthering this cause, and we need to provide consistent, usable access to information for the community of people doing this very work. The wide recognition of FAO as a neutral international centre of excellence in agriculture positions it perfectly to lead in the growth and improvement of knowledge representation systems in the agricultural domain.

Above discussed Semantic Web applications could contribute to this mission. The need for improved information management mechanisms within the various knowledge domains of this organization is therefore evident.

The Agricultural Ontology Service (AOS) Project evolved from this motivation and has been initiated to act as a reference tool for ontology creation to enable the transfer of the agricultural information domain towards the Semantic Web. The goals of the AOS are to increase the efficiency and consistency of describing and relating multilingual agricultural resources, to decrease the random nature and increase the functionality for accessing these resources and to enable sharing of common descriptions, definitions and relations within the agricultural community. To achieve these goals the AOS assists community partners in the creation of ontologies and related activities. The project, which will be presented in this document, serves as a prototype within the AOS framework and shall serve as a reference to further activities. A comprehensive and reusable methodology, which can be applied to any other domain, is to be evaluated by this prototype. A multilingual, extendable model for the representation of domain ontologies builds the core baseline of this methodology and will be presented in the following section.

3. The ontology: Modelling and representation

In the context of the AOS, an ontology is a system of terms, the definition of these terms and the specification of relationships between the terms. It extends the approach of classical thesauri by providing the opportunity of creating an infinite number of different semantic relationships. For an overview about different types of ontologies, refer to (Guarino 1998). The following gives a detailed description of the modelling approach used for our representation of the prototype ontology:

Semantic robustness towards representational changes, as well as multilingualism, are crucial for the development of this domain ontology (see section 1.1). Therefore, we distinguish between terms, and the concepts these terms represent. Whereas terms might change, and are different in each language, the semantic meaning and interpretation of the terms' abstract concept stays the same[2]. In the presented modelling approach, a concept's term representations are called Lexical Entries. These Lexical Entries are limitless and may be characterized as labels, synonyms or word stems. Furthermore, each Lexical Entry has at least two attributes: the concept it refers to and its language. Lastly, relationships between concepts can be established, annotated by the same lexical entries. This approach can be described as a two layered model, in which the semantic layer of the ontology is totally independent from its representation layer and hence, robustness against changes can be achieved.

Ontologies can be represented in different representation languages. (Palmer 2001) gives a brief overview about these languages and provides further information. RDFS[3], the language that was chosen to

[2] This holds in most cases. There are however cases, where a concept does exist in one culture, even though there is not adequate concept in another one. This is however more evident in humanity domains, since concepts there are richer and less well defined. The project environment here is rather technical and hence chances for this can be neglected.

[3] http://www.w3.org/TR/rdf-schema/#intro.

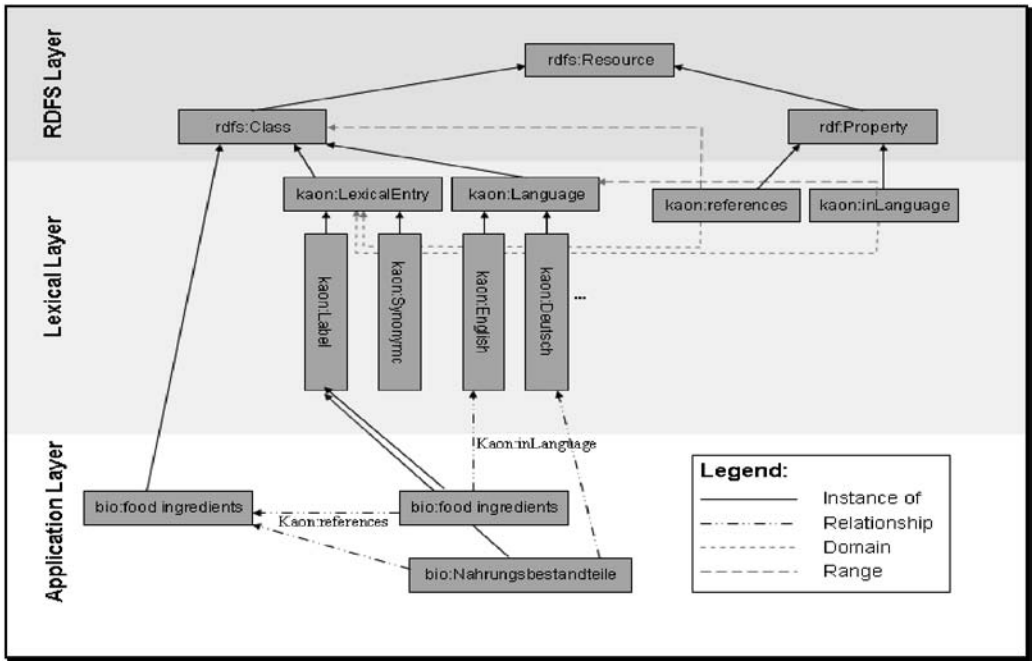

Figure 2. Layered RDFS model multilingual ontology representation

be used within the AOS framework, is used to define vocabularies of resources and relationships amongst them. Resources can be documents, web pages or parts of them, basically anything, which can be referenced by a URI[4]. RDFS provides a basic set of modelling primitives, which can be easily extended by users to include domain specific semantics in terms of relationships among concepts. Furthermore RDFS models are exchanged via XML and therefore provide interoperability between communities. Although still under development, RDFS evolves to serve as a standard representation in the context of the Semantic Web. For a detailed discussion about modelling ontologies in RDFS, refer to (Staab et al. 2000a).

Figure 2 gives an overview of the above-discussed layered modelling approach in RDFS. The top layer represents an extract of the basic layer provided by the RDFS language. The lexical layer creates the needed abstraction of lexical and language representation from conceptual domain semantics. The lowest layer finally constitutes the domain. The most generic class in RDFS is rdfs:Resource[5], from which every other class is derived An rdfs:Class can be instantiated to define domain specific concepts. Lexical Entries are separate classes which can be instantiated and attached to concepts using the properties kaon:references and kaon:inLanguage. Each property has a domain and a range, which determine the source and the target of the relationship respectively. In that way, an infinite number of lexical entries can be instantiated and related to domain concepts and different languages. If a representation of a concept in terms of its lexical entry changes, the semantics of the ontology are not affected, since it still refers to the same concept. Furthermore, additional domain properties can be derived from rdf:Property in the application layer to relate the domain concepts and build the semantic network.

This generic, multilingual ontology model establishes the basis for our engineering methodology framework, which will be presented in the following section.

4. The methodological framework

Until now, few domain-independent methodological approaches have been reported for building ontologies. Most of these are mainly overall lifecycle models providing a more generic framework for the ontology creation process, but giving little support for the actual task of building the ontology. A comparative study of ontology building methodologies from scratch can be found in (Fernandez 1999). The METHONTOLOGY methodology, as described in (Fernandez et al. 1998) fits our project approach best, since it proposes an evolving prototyping life cycle composed of development oriented activities (requirements specification, conceptualization of domain knowledge, formalization of the conceptual

[4] Uniform Resource Identifier. See also http://www.w3.org/Addressing.

[5] The prefixes rdfs: , rdf: , kaon: , bio: represent XML namespaces and are to uniquely identify each resource. Refer to (RDFSchema 2002) to learn more about RDFS and namespaces.

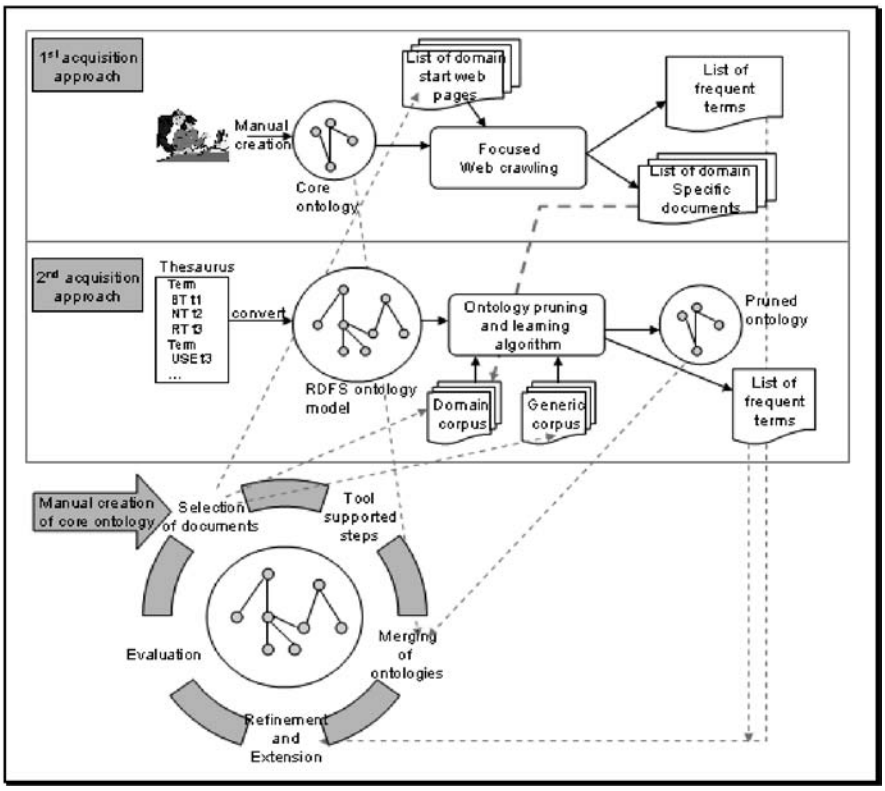

Figure 3. Comprehensive framework for creation of domain ontologies

model in a formal language, implementation of the formal model and maintenance of implemented ontologies), support oriented activities (knowledge acquisition, documentation, evaluation, integration of other ontologies) and project management activities. Since this has been done elsewhere, the framework presented in this paper will not propose another life cycle model. Rather, it will depict the development oriented activities within the above methodology and provide a more specific methodology for this part. More specific methodologies, especially for supporting the creation process sparsely exist so far. (Guarino et al.) provide a set of methodologies for ontology-driven conceptual analysis. An overview of these methodologies can be accessed through his web site. The methodology presented here focuses on the actual acquisition and development part of the ontology and describes a comprehensive, reusable and semi automatically-supported framework, which can be embedded in other lifecycle models. Figure 3 shows a graphical overview of the overall framework.

The domain ontology is built using two different knowledge acquisition approaches, which will be described in detail in the following sections. The top of the picture describes these two paths. In the lower part of the picture the cyclic evolvement of the domain ontology to be built is shown. The grey dashed arrows show how outputs of certain processes steps are used as inputs of other steps. Section 5, where the application of this framework to the biosecurity prototype is presented, will present each single process step and its application to the prototype project.

5. The biosecurity ontology project

5.1 Acquisition approach 1: Creation of the core ontology

In the first acquisition approach, a small core ontology with the most important domain concepts and their relationships is created from scratch. This stage is basically comprised of the first three steps of the METHONTOLOGY development activities (as described in section 4):

First the goal of the ontology is specified (as outlined in section 1.1 and in section 2). In a second step, subject specialists accomplish the conceptualization of the core model. The Codex Alimentarius, which serves as a reference for food standards in food safety biosecurity, has been chosen here for extracting basic domain concepts. In further brainstorming sessions, relationships between the chosen concepts and additional concepts are created. The concepts and relationships are further assessed using criteria including clarity, ambiguity, unity and rigidity. A detailed discussion of criteria for ontology-driven conceptual analysis is given in (Welty 2001).

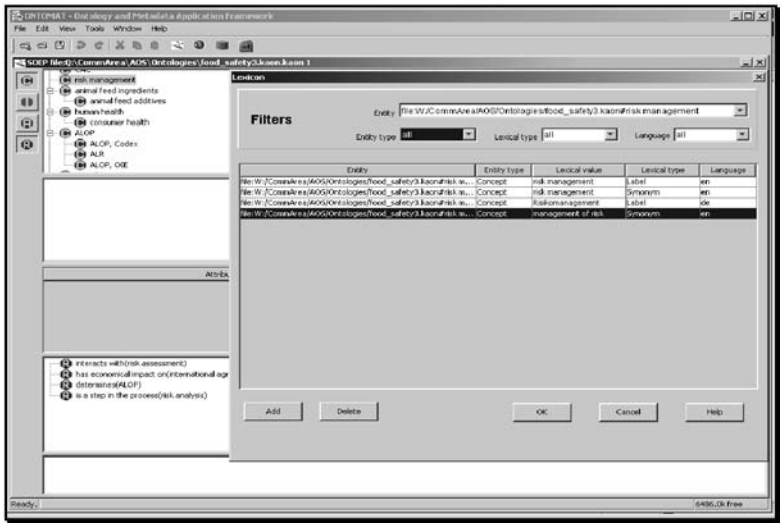

Figure 4. Screenshot of the ontology editor SOEP

In the biosecurity project, this initial step created a core ontology with 67 concepts and 91 relationships connecting these concepts, equalling an average rate of 1.36 relationships per concept.

Finally the developed core ontology is formalized in the formal RDFS language. This can be accomplished using the RDFS compatible ontology editor SOEP[6] of the KAON[7] tool environment. The editor has an easy-to-use graphical user interface, which allows the creation of the concepts, their relationships and their lexical entries. Figure 4 shows a screenshot of the resulting core ontology in the editor. On the upper left, concepts and their hierarchical subclass relations are shown. On the lower left, one can see the domain specific relationships between a marked concept and other concepts. The additional window on the right side shows the lexical layer of the ontology. This clearly illustrates that the entities (in this case the concept 'risk management') are represented uniquely by a URI, therefore unambiguous, and a concepts lexical entries are all independently associated with this URI.

In the following acquisition stage, the core ontology is fed into a Focused Web Crawler, another tool of the KAON environment. The Crawler takes a set of start URLs and domain ontology. It then crawls the web in search of other domain specific documents based on a large set of user specified parameters. The outcome this process creates consists of a rated list of found domain specific documents and links as well as a list of most frequent terms found on these documents. A list with 264 domain-relevant web pages and a list with 36 frequent terms have been output by the crawler in our prototype project. The list of keywords can later be used to extend the core ontology. The document list can be used as input in the second ontology acquisition approach, which will be described in the following section.

5.2 Acquisition approach 2: Deriving a domain ontology from a thesaurus

The second approach towards ontology acquisition takes a well-established thesaurus as starting point. Here, AGROVOC[8], a multilingual agricultural thesaurus consisting of almost 30,000 keywords developed by the FAO, is assumed to contain domain descriptors. A thesaurus like AGROVOC consists of descriptive keywords linked by a basic set of relationships. The keywords are descriptive in terms of the domain in which they are used. The relationships may either describe a hierarchical relation or an inter-hierarchical relation. For example, 'Broader Term' and 'Narrower Term' are used for the former and 'Related Term' and 'Use' for the latter. The 'Use' relationship indicates that another term should be used for description instead of this one.

The process begins by representing the thesaurus in an adequate format, where an ontology can be derived from. As discussed above, RDFS is chosen as the representation language. Then, as done in the biosecurity ontology, all terms of the thesaurus are converted to classes (concepts)[9]. The Broader and Narrower Term relationships are used to form the hierarchical class-subclass structure, which constitutes the basic taxonomy of the ontology. Finally the Related Term and Use relationships are represented as properties of the classes and form an initial set of non-hierarchical relationships. This approach extends the basic RDFS language by creating new, layered meta-properties, which can be instantiated in

[6] Simple Ontology and Metadata Editor Plugin.
[7] Karlsruhe Ontology and Semantic Web Tool Suite.
[8] http://www.fao.org/agrovoc
[9] In this paper, classes and concepts are synonymous, where class refers to the RDFS representation of the concept in an ontology.

```
...
<rdf:Property rdf:ID="rt">
        <rdfs:domain rdf:resource="&kaon;Root"/>
        <rdfs:range rdf:resource="&kaon;Root"/>
</rdf:Property>
....
<rdfs:Class rdf:ID="7">
        <rdfs:subClassOf rdf:resource="#1172" />
        <rt rdf:resource="#3471" />
</rdfs:Class>
...
```

Figure 5. Extract of RDFS modelling of the AGROVOC thesaurus, using meta properties

the domain classes. The modelling is done analogously to the above described language layer. Figure 5 gives an example representation of the Related Term definition and a class using this relationship in RDFS. Here the concept with the identifier 7 is a sub class of concept 1172 and is related to the concept with the identifier 3471. Lexical labels for representation in different languages are attached to these concepts and relations as discussed before.

The converted thesaurus still has to be trimmed to the specific domain. An ontology pruner is used to accomplish this task. In order to prune the thesaurus structure to extract a domain-specific ontological structure, two sets of documents are needed: a domain specific set, descriptive for the domain of the goal ontology to be built, and a generic set, containing a representative set of generic, unspecific terms. This step can partly be done before the tool supported steps and therefore appears on top of the cyclic process in Figure 3. The domain documents have to be carefully chosen by subject specialists. The output of the process obviously correlates with the descriptiveness, preciseness and richness (in means of specific domain term usage) of the domain document set. The document list, which is the outcome of the web crawling process, can serve as a good source. Publicly available reference corpora and newspaper archives serve as sources for the generic corpus. In addition, sets of related, but different, subject domains may also be used. This could increase the chances of retrieving only very specific concepts, since the terms' frequencies of the domain corpus are measured against those of the generic corpus. However, the whole process is a highly heuristic approach and further experiments are needed to establish a significant document set quality measure.

In our case, a set of six domain specific documents (mainly excerpts of the Codex Alimentarius, as well as documents about food safety and risk assessment) has been chosen and another eight documents have been taken from the list of the crawling process. The generic document set has been compiled using news web pages, as well as pages from the animal feed domain, another research area within the FAO.

In order to prune domain unspecific concepts, concept frequencies are determined from both domain-specific and generic documents. All concept frequencies are propagated along the taxonomy to their super concepts by summing the frequencies of sub concepts. The frequencies of the concepts in the domain corpus are then compared with those of the same concepts in the generic corpus using pruning criteria. Only the concepts, which are significantly more frequent in the domain corpus, remain in the ontology, the others are discarded. Moreover the frequencies of all terms occurring in the domain documents can be compared against all the terms that occur in the generic corpus resulting in a list of terms, likely to be significant for the domain corpus. Refer to (Volz 2000) for a detailed discussion on ontology acquisition using text mining procedures and to (Kietz 2000) for a similar application of extracting a domain ontology.

The result of the second ontology acquisition approach is a pruned ontological structure derived from the original thesaurus, containing only the domain specific terms. It also produces a list of likely domain-specific terms, which can serve as possible candidates for the ontology refinement process.

Here, an ontological structure with 504 concepts could be extracted from the AGROVOC thesaurus with a taxonomic depth of five. A list of 1632 frequent terms has been produced from the domain document set.

5.3 Ontology merging

The above acquisition steps have created two ontologies, the manually created core ontology and the derived ontology, using thesaurus terms. These have to be assembled into a single ontology. Ontology merging is still more of an art than a well-defined and established process. (Gangemi et al.) describe a methodology for ontology merging and integration in the Fishery Domain. Besides the editor environment, computer support for this process is not available and therefore needs extensive subject specialist assessment.

From the pruned ontological structure of the AGROVOC thesaurus, 23 concepts and 13 instances have been extracted to extend the core ontology in our case. Hence, almost 10% of the automatically extracted knowledge could be used in the first instance. More terms might serve as candidates in further refinement steps.

5.4 Ontology Refinements and Extension

The second result produced by the acquisition steps is a list with frequent domain terms serving as possible candidate concepts or relationships for extending the ontology. These terms have to be assessed by subject specialists and checked for relevance to the ontology. The same principles and methodologies, as in the creation process of the core ontology, apply to this session. In our case, 12 concepts were directly taken from the lists of frequent

keywords to extend the ontology. A set of 12 new unique relationships has been defined, resulting in 92 relations interlinking and integrating the newly created concepts. These have been applied to assemble the final prototype ontology consisting of 102 concepts, 12 instances and 183 relationships among the concepts. This corresponds to an average rate of 1.79 relationships per concept, representing a higher density than in the core ontology.

The resulting ontology is now subject to more extensive evaluation and testing by a broader audience. The presentation of the ontology in a multilingual portal, which will be presented in the next section, can help in the evaluation process. However, extensive testing and evaluation cannot be done effectively until real applications utilize the semantic power of the ontology. This will be addressed in the last section, where an outlook on further work and future uses will be given.

5.5 Presentation in Multilingual Portal

The domain ontology can be extended to represent the concepts in multiple languages. The translation process has to be done manually, since current translation tools show rather inferior performance and are also quite unlikely to be applicable to specific domains like the biosecurity portal. With our ontology model introduced in Section 3, this task can easily be achieved by simply attaching further lexical entries to the concepts of the newly created ontology. In the project presented here, this step has been omitted since it is not of importance to prototype versions. Finally, KAON PORTAL, a web-based portal to present RDFS based ontologies, can be used to present the ontology, making it available and browseable to the target community. Figure 6 shows a screenshot of the top concept layer of the prototype Biosecurity Ontology. The display can be switched to different languages, including Arabic and Chinese.

This portal could now be extended to actually link to a domain document base and the ontology could be used to perform semantically extended search opportunities.

6. Outlook: Future uses of the ontology (implementation of the semantic web)

In this section, an outlook on future work within this project and follow-up projects in context of the AOS framework will be given. As previously discussed, a domain ontology, which can be developed applying the above framework, only sets the basis for efficient information management and retrieval. Applications, using this background knowledge are still rare and further investigation is required. This section sketches a likely scenario for ontology use in the discussed domain and outlines some already existing sample applications and their possible implications for the AOS project.

6.1. Facilitation of better search and information retrieval

Using the ontology to extend currently performed keyword search, is the most direct application. Ontology based support could be given at two stages of the search query process: before the actual execution of the query and/or after retrieval of the results. Figure 7 shows these two semantically enhanced search features. The left side shows a scenario, in which the ontology assists the user by providing an easy way to extend or refine her search. The ontology enabled search application processes on the initial search term. It then queries the ontology to retrieve the

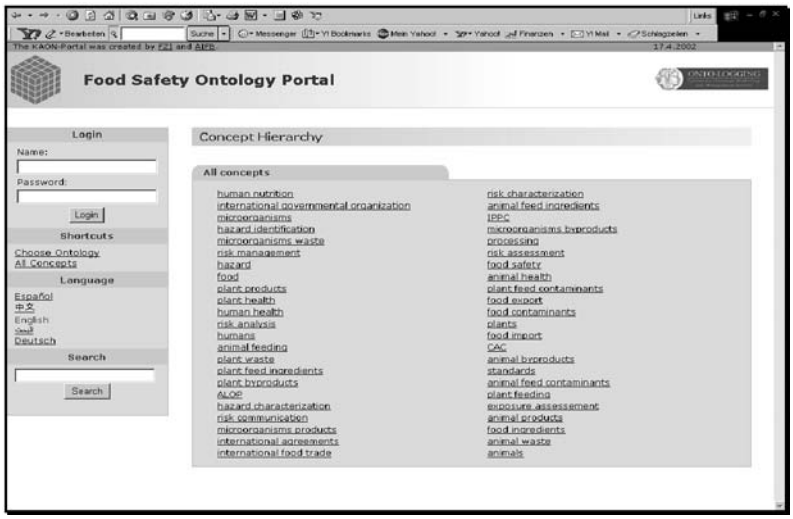

Figure 6. Screenshot of multilingual, web based ontology browser

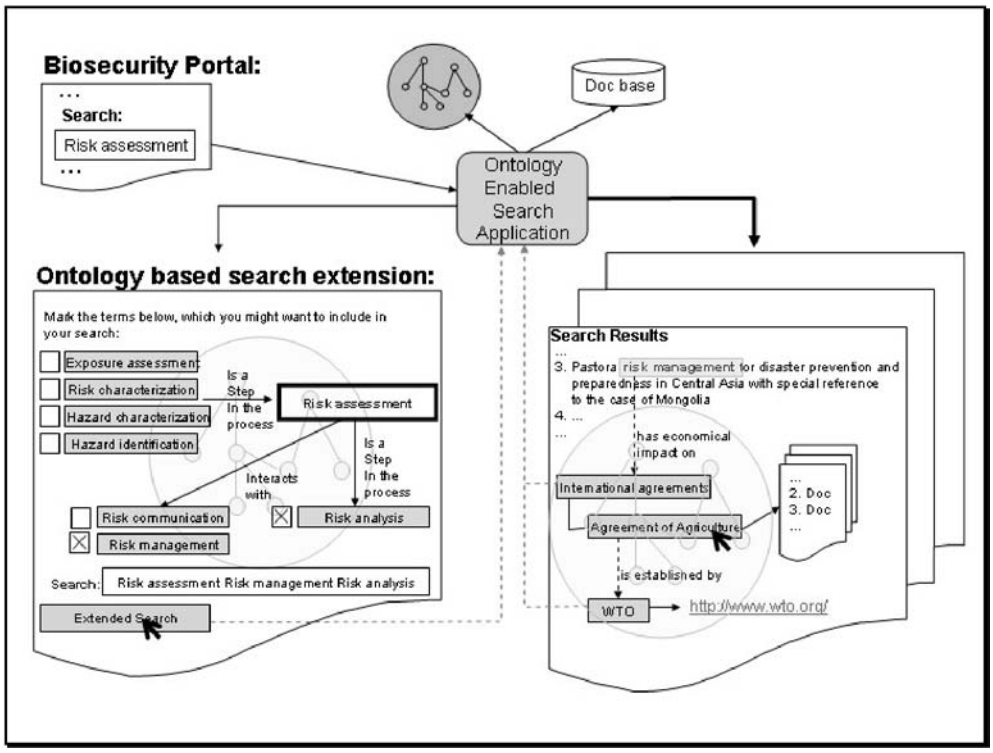

Figure 7. Ontology based search extension and semantically structured result display

semantic context of the search, and returns the results back to the user, giving her the possibility to extend or specify the query. The interlinked grey boxes show the conceptual neighbourhood of the search term in the Biosecurity Ontology prototype. The extended query is passed back to the application, which now searches the document base. Once again, the semantic context within the ontology can be used in order to provide the user with related results which might be of interest. The picture on the right shows an excerpt of retrieved search results. The user is provided with additional links or documents, which are related to neighbouring concepts of the initial search term. This shows how domain ontologies can be useful in knowledge discovery and providing domain relevant, semantic links among search results.

These features have yet to be implemented and evaluated in future project work. Hence, usability has not been proven at this stage.

A commercially available tool providing similar functionality (like automatic keyword search extensions and structured, enhanced result representation) is the Semantic Miner from Ontoprise[10].

In the above discussed solution, the annotation of the documents does not change and the same document bases are accessed. A further step would be, to actually annotate the documents of the domain of interest with the semantic information of the ontology. With semantic annotation, not only support in search term compilation and semantic structuring of search results can be given, but documents and their annotated content can now be interlinked semantically to provide enhanced knowledge discovery. Refer to (Staab et al. 2000b) for a detailed discussion of semantic annotation.

6.2. Semiautomatic, ontology based text classification

Text classification is a time-consuming task, which is typically performed manually. However, the vast amount of information on the internet makes it impossible to continue using this approach for arbitrary web documents in the future. Statistical classifiers exist and have shown quite good results using standard texts, which all follow certain patterns. A good overview about methods and evaluations is given in (Aas 1999). However, none of the methods can so far replace human classifiers, since they all lack the specialist's semantic knowledge of the domain in which the document has to be classified. Little research has been done in integrating ontological background knowledge into classical text classification methods. One attempt[11] used the freely available dictionary WORDNET[12] to serve as background knowledge for text classification with support vector

[10] http://www.ontoprise.de/com/download/semminer_iswc_submission.pdf.

[11] A research study done at the University of Karlsruhe in 2002; refer to (Pache 2002) for details.

[12] http://www.cogsci.princeton.edu/~wn/.

machines. The classifier used the News20-document-set for evaluation purposes and showed good performance. This work can now be expanded, and WORDNET can be replaced with a domain ontology, such as the Biosecurity Ontology, to evaluate the classifier against arbitrary web documents. An automatic indexing approach like this could then be used in combination with Dublin Core elements to index web pages for Semantic Web purposes.

7. Summary

We have presented a new approach towards domain ontology creation. The introduced framework provides a generic, reusable methodology, which can be reapplied to create domain ontologies in various fields of interest. The prototype project which has been presented in this paper showed the applicability of the methods in the biosecurity domain. We introduced a generic layered ontology modelling approach that can be used to describe any possible domain of interest. Multilingual aspects have been addressed to solve the problems of portability, usage and representation of semantic knowledge in different languages. The overall framework, we described in Section 4 and 5, provides a comprehensive methodology for domain ontology creation and is not bound to any domain specific input. Used thesauri, document sets and core ontologies can easily be replaced by equivalents from other domains. Moreover, as the open source applications are all Java-based, the used toolset providing the semiautomatic support is extremely adaptable to different needs. Obviously, the whole approach is completely portable and reusable in other domains.

We concluded our presentation, giving an outlook on further work to be done in the field of domain ontology usage. Example scenarios and applications have been addressed, giving an outlook on possible implementations of the Semantic Web: The initial motivation for building ontologies.

Acknowledgements: This project has been done in close collaboration with the AIFB[13] institute of the University of Karlsruhe (TH) in Germany. All tool supported steps have been carried out, using the freely available KAON environment, developed at this institute. We would like to express our gratitude to the KAON group (KAON) for their technical expertise in this subject. We particularly thank Raphael Volz for his sound direction, technical guidance and supervision throughout the project. We also gratefully recognize the programming support of Boris Motik, which facilitated the adaptation of the tool set.

[13] Institut für Angewandte Informatik und Formale Beschreibungsverfahren, Universität Karlsruhe (TH), Karlsruhe, Germany.

References

Aas K., Eikvil L., Text categorisation: A survey. Technical report, Norwegian Computing Center, June 1999.

AOSProposal 2002. http://www.fao.org/agris/aos/Documents/AOS_Draftproposal.htm. June 2002.

Fernandez M., Blazquez M., Garcia-Pinar J.M., Gomez-Perez A., 1998. Building Ontologies at the Knowledge Level using the Ontology Design Environment.

Fernandez M., Gomez-Perez A., Pazos Sierra A., Pazos Sierra J., 1999. Building a Chemical Ontology Using METHONTOLOGY and the Ontology Design Environment. IEEE Expert (Intelligent Systems and Their Applications), 14(1): 37-46.

Gangemi et al., A Formal Ontological Framework for Semantic Interoperability in the Fishery Domain. February 2002.

Guarino N., Formal Ontology and Information Systems. In: N. Guarino, editor, Proceedings of the 1st International Conference on Formal Ontologies in Information Systems, FOIS'98, Trento, Italy, pages 3-15. IOS Press, June 1998.

http://www.ladseb.pd.cnr.it/infor/ontology/methodology.html. June 2002.

Bozsak E., Ehrig M., Handschuh S., Hotho A., Mädche A., Motik B., Oberle D., Schmitz C., Staab S., Stojanovic L., Stojanovic N., Studer R., Stumme G., Sure Y., Tane J., Volz R., Zacharias V., KAON – Towards a large scale Semantic Web. In: Proceedings of EC-Web 2002. Aix-en-Provence, France, September 2-6, 2002. LNCS, Springer, 2002.

Kietz J.-U., Volz R., Maedche A., Extracting a Domain-Specific Ontology from a Corporate Intranet. Proc of the 2nd Learning Language in Logic (LLL) Workshop, Lissabon. September 2000.

Pache G., 2002. Textklassifikation mit Support-Vektor-Maschinen unter Zuhilfenahme von Hintergrundwissen. Studienarbeit. AIFB Universität Karlsruhe(TH), Karlsruhe, Germany. April 2002.

Palmer S., 2001. The Semantic Web: An Introduction. http://infomesh.net/2001/swintro.

RDFSchema 2002.
http://www.w3.org/TR/rdf-schema/#intro. June 2002.

Staab S., Erdmann M., Mdche A., Decker S., An Extensible Approach for Modeling Ontologies in RDF(S). In: Proceedings of ECDL 2000 Workshop on the Semantic Web, 11-22, 2000.

Staab S., Mädche A., Handschuh S., An Annotation Framework for the Semantic Web. In: S. Ishizaki (ed.), Proc. of The First International Workshop on MultiMedia Annotation. January, 30-31, 2001. Tokyo, Japan.

Volz R., Akquisition von Ontologien mit Text-Mining-Verfahren. Technical Report 27, Rentenanstalt/Swiss Life, CC/ITRD, CH-8022 Zürich, Switzerland, ISSN 1424-4691.

Welty C., Guarino N., 2001. Supporting Ontological Analysis of Taxonomic Relationships. Data and Knowledge Engineering 39(1), pp. 51-74.