

Building an Allergens Ontology and Maintaining it using Machine Learning Techniques

Alexandros G. Valarakos et al.

Software and Knowledge Engineering Laboratory,
Institute of Informatics and Telecommunications,
National Centre for Scientific Research (NCSR) “Demokritos”,
153 10 Ag. Paraskevi, Athens, Greece

Department of Information and Telecommunication Systems Engineering,
School of Sciences, University of the Aegean,
83200 Karlovassi, Samos, Greece

Abstract

Ontologies are widely used for formalizing and organizing the knowledge of a particular domain of interest. This facilitates knowledge sharing and re-use by both people and systems. Ontologies are becoming increasingly important in the biomedical domain since they enable knowledge sharing in a formal, homogeneous and unambiguous way. Knowledge in a rapidly growing field such as biomedicine is usually evolving and therefore an ontology maintenance process is required to keep ontological knowledge up-to-date. This work presents our methodology for building a formally defined ontology, maintaining it exploiting machine learning techniques and domain specific corpora, and evaluating it using a well defined experimental setting. The application of this methodology in the allergen domain is then discussed in detail presenting the ontology built, the specific techniques used and the evaluation settings.

This is a pre-print version

1 Introduction

The problem of efficient access to the required information is faced by all professionals nowadays due to the growing volume of information constantly

flowing over the World Wide Web and other media, in various formats and languages. This problem is also crucial for physicians and researchers in medicine and biology, who need efficient access to up-to-date information according to their interests and needs. MEDLINE contains more than 12 million citations and every week approximately 2000 are added. Researchers need to digest a large number of publications relevant to a specific domain and combine the information extracted from them. Therefore, the use of intelligent techniques for automating the information retrieval and extraction task is a major need for rapidly growing fields such as biomedicine. The use of ontologies which describe and formalize the terminology and knowledge for a domain is an essential element of such intelligent techniques. For example, suppose that several different Web sites contain medical information. If these web sites share a common understanding of the structure of information through the same underlying ontologies of the terms they use, then information retrieval and extraction systems will be able to locate and aggregate information from all these sites. The significance of ontologies is shown by the continuous growth in the number of ontologies being integrated in industrial and academic applications, especially in fields supporting semantics-based search, interoperability support, semantic web applications and others [1].

Ontology construction in general involves the following steps: selection of concepts to be included in the ontology, specification of concepts' attributes and relations between concepts, and population of the ontology with instances filling their concepts' properties¹. During the last years, several methodologies and tools for building ontologies have been presented. In this work, we studied existing methodologies and design criteria for ontology building and we examined several existing ontologies and databases for the domain of allergens in biomedicine. The problems we identified in existing allergens ontologies and databases motivated us to design and implement a formally defined ontology for allergens by exploiting and refining existing methodologies and tools. We also gave special emphasis to the problem of ontology maintenance due to relevant deficiencies identified in existing ontologies and databases for allergens. For this purpose, we examined and refined our methodology for the semi-automatic maintenance of ontologies [2] in the domain of allergens. The outcome of the above effort is a methodology for the whole process of ontology development: designing and implementing a formally defined ontology, maintaining it using a machine learning based approach, and evaluating the maintenance process.

¹Property refers to concept's attributes and relations.

Section 2 presents related work in ontology building and maintenance and provides information on existing allergen specific resources. It focuses on the problems of such resources and the need for building a formally defined ontology which can be semi-automatically maintained. Section 3 presents the major stages of our methodology for ontology building, maintenance and evaluation. The application of each of these stages in the allergens domain is discussed in detail in sections 4 (building the allergens ontology), 5 (populating and enriching the allergens ontology exploiting a corpus of PubMed² abstracts on allergens), and 6 (evaluating the ontology maintenance process and discussing the evaluation results). Finally section 7 concludes and presents our future plans.

2 Related Work

Ontology building is a complex task and requires special attention in order to build a useful and machine exploitable ontology [3]. A few methodologies for building an ontology from scratch exist and propose a concrete number of stages towards its development (see relevant surveys: [4], [5] and [6]):

- *Specification*: the purpose of the ontology is specified in order to restrict the various conceptual models that can be used for modeling (conceptualizing) the domain.
- *Conceptualization*: terms that represent concepts, their attributes and relations are enumerated in order to form the conceptual description of the ontology.
- *Formalization*: the conceptual description of the previous stage is transformed into a formal model. Formal definition of concepts is performed through axioms that restrict the possible interpretations for the meaning of these concepts, as well as through relations that organize the concepts such as is-a or part-of relations.
- *Implementation*: the formalized ontology is implemented using a knowledge representation language.
- *Maintenance*: the implemented ontology is periodically corrected and updated. This is a crucial stage for the exploitation of an ontology since ontologies used in practical applications need to evolve taking

²<http://www.ncbi.nlm.nih.gov/entrez/query.fcgi>

into account the changes occurring in the domain knowledge they represent, i.e. new instances or variants of existing instances are introduced, new concepts may have to be added, etc.

The methodologies mentioned are associated with general guidelines for each stage, such as specify ontology purpose, identify relevant terms, make informal and formal text analysis, interview domain experts etc. However, they lack information on how to accomplish these guidelines, in terms of the specific actions and decisions that must be performed in each stage. In section 3, we discuss the design criteria that must be taken into account in the design (specification, conceptualization, formalization) and implementation stages and we present some examples from their application in practice.

Concerning the maintenance stage, this involves adding new instances (ontology population), by filling concepts' properties, as well as adding new concepts and concept's properties (ontology enrichment). A representative example of ontology population work is presented in Craven et al. [7]. They state the need for constructing and maintaining knowledge bases with information coming from the Web and stress the need of formally storing information in knowledge bases which results in a more effective and intelligent information retrieval exploiting knowledge-based inference capabilities. Given a domain ontology and a set of manually provided training examples, i.e instances that belong to particular concepts, their system learns to extract new ontology instances and properties from the Web. A recent work on ontology population is the autonomous KnowItAll system [8] that incrementally extracts information from the web in an unsupervised way given only an initial ontology of a particular knowledge representation formalism. The system starts by instantiating 8 domain-independent generic extraction patterns, inspired from Hearst work [9], using the labels of the concepts and relations of the initial ontology, producing a set of extraction rules. [10, 11] are also presenting relevant work in ontology population focusing mainly in the extraction of instances from textual corpora using information extraction systems. In these efforts, the training examples for the extraction systems are provided by manually annotating a corpus, whereas our approach relies on the automatic creation of the training examples exploiting an initial version of the domain ontology with limited size and domain-independent rules [2]. In Ciravegna et al [12], learning is enforced by integrating information from various structured sources, e.g. databases and digital libraries. A rule-based approach is adapted by Kiryakov et al. [13] who tackle the problem as a named entity recognition task combining linguistic analysis and manually crafted rules to populate an ontology that contains many gen-

erally used entity types such as persons, companies etc. Although this work intends to identify domain-independent name entities, the use of manually crafted rules for their exploitation in web pages makes harder the migration to other applications.

Efforts in ontology enrichment mainly focus on associating concepts with an already known relation [14, 15, 16, 17] or adding a new concept along with a new discovered relation [18, 19, 20, 21, 22] that associates the new concept with an already known concept of the initial ontology. These efforts extract knowledge from textual corpora. The main difference between these two approaches is that in the former the relation is known from the beginning, hence lexico-syntactic patterns can be built or learned by examples of it. Whereas, in the later one, new concepts are discovered through an implied relation that is subjectively defined (various similarity measures are being used) exploiting features from the context of the concept's lexicalization in the corpus. Our work belongs to the former approach as we enrich ontology with a lexico-semantic relation in an unsupervised way. It is the first time that such a relation is being addressed in the context of ontology enrichment task.

Ontologies represent the solution to the semantic and structure heterogeneity that appears in database schemata since they are able to provide a shareable, consistent and formal description of the information source semantics [3]. Various biomedical communities have created several ontologies in order to address the interoperability problem [23] between the various database applications [24] or to provide a common vocabulary and semantics [25, 26, 27, 28]. In the context of the OBO³ (Open Biological Ontologies) project, several biomedical ontologies or controlled vocabularies can be found. However, the knowledge in a domain ontology is usually evolving especially in dynamic domains such as most of the domains in biomedicine. For example, new allergen names or variants of existing ones appear frequently in the literature, following or not the scientific nomenclature. Considering that the regular update of a domain ontology is crucial for its reliability and quality, the process of ontology maintenance is a necessity in the area of biomedicine.

Various allergen databases and lists exist most of which are freely available on the web. Their schemata are more or less similar, concentrating mostly to allergen's name, the species it occurs in and the protein associated with, along with its links to GenBank and SwissProt through their accession number. The Allergen Nomenclature Sub-Committee of the In-

³<http://obo.sourceforge.net>

ternational Union of Immunological Societies⁴ (IUIS) maintains a list of ‘certified’ protein allergens consisting of a list of allergens, isoallergens and variants. References to GenBank or SWISS-PROT accession numbers are provided for some of the allergens. SWISS-PROT provides an organized index⁵ of allergen sequences that contains allergen names, accession numbers and links to SWISS-PROT entries. The Biotechnology Information for Food Safety⁶ (BIFS) contains a non redundant allergen protein list constructed in order to carry out sequence comparisons for the assessment of potential allergenicity of proteins used in foods. The BIFS tables contain lists of sequences published in GenPept, SWISS-PROT and PIR, providing for each entry - when available - origin species, protein name, allergen nomenclature, literature references, general notes and the available GenPept, SWISS-PROT and PIR accession numbers. The PROTALL database⁷ at the Institute of Food Research of the United Kingdom focuses on various types of content related to food allergy. A PROTALL entry contains general allergen information (name, designation, accession numbers and links in major databases, structural information), specific properties of the allergen (molecular weight, structure, epitopes, stability, cross-reactivity, protein function, purification information and references) and clinical data (oral provocation information, skin prick test, IgE determination assay, clinical history and references), as an effort to bring together information from specialists such as clinicians, food scientists and plant biologists. ALLALLERGY⁸ database contains protein as well as chemical allergens and can be queried using the name or the allergen category (such as fish, fruit, pollens) and returns the background information, the function of the substance and the adverse reactions (personal communication). The Structural⁹ Database of Allergenic Proteins contains structural and epitope-related information on allergens. Finally, ALLERbase¹⁰ is a searchable protein allergen database with links to external accessions (PIR and GenBank), but not up-to-date. For further details on available allergen databases and data sources, see the review article [29]. The main problem of these schemata is related to the differences occurring in the meaning of their categories (semantic heterogeneity) as well as their structure (structure heterogeneity). For example, some schemata use the

⁴www.allergen.org/List.htm

⁵<http://www.expasy.ch/cgi-bin/lists?allergen.txt>

⁶<http://www.iit.edu/~sgendel/fa.htm>

⁷<http://www.ifr.bbsrc.ac.uk/Protal>

⁸<http://www.allallergy.net/>

⁹<http://fermi.utmb.edu/SDAP>

¹⁰<http://www.dadamo.com/allerbase/allerbase.cgi>

term ‘trivial name’ to refer to the allergen’s common name and the term ‘description’ to refer to the protein associated with, in contrast to others that use the term ‘common name’ and ‘biochemical id’, respectively. Other databases label the allergen source as ‘species’ and fill it either by the organism ‘scientific name’ (*Olea europea*) in which the allergen occurs or its ‘common name’ (Olive tree). Also, some of the databases provide unstructured information (ALLALLERGY). Moreover, these schemata are highly ambiguous because they do not provide rigor definitions of the vocabulary uses, e.g. what is the meaning of source, should be filled with allergen sources or proteins, etc. And finally, many of them are not updated regularly, thus being out-of-date.

The above problems motivated us to design and implement a formally defined ontology for the allergen domain exploiting existing ontologies and databases, according to a set of well defined design criteria (see section 3.1). In addition, these problems motivated us to examine the application in the allergen domain of our methodology for the semi-automatic maintenance of ontologies using machine learning techniques (see section 3.2). The application in the allergen domain enabled us to refine and extend our methodology for ontology maintenance, presenting at the end a methodology for the whole ontology development process, from the design and implementation of a formally defined ontology to its maintenance and evaluation. This methodology is presented in the next section whereas its application in the allergen domain is described in detail in sections 4-6.

3 Designing and Maintaining a Domain Ontology

This section presents the major principles we use to design and build a domain ontology in the context of the widely accepted stages of the ontology building methodology presented in Section 2. It then describes the major steps of the proposed methodology for automating the ontology maintenance process and evaluating it in a new application domain. The application of all 3 stages of the proposed methodology (design and implementing, maintaining, evaluating) in the allergens domain is presented in detail in sections 4-6 respectively.

3.1 Ontology Design

As noted in Section 2, the stages through which an ontology is usually built are: specification, conceptualization, formalization, implementation and maintenance. This sub-section discusses the design criteria we followed

in the first four stages. The methodology we propose for automating the ontology maintenance stage is presented in the following sub-section. According to Gruber [30, 31], the general design criteria for ontology engineering are the following:

- *Clarity*: context-independent, unambiguous, precise definitions. This is directly related to the conceptualization stage where the conceptual model of the ontology to be built is described (i.e. ontology concepts, their attributes and relations). This criterion is also related to the formalization stage since formal models provide the means to define concepts using necessary and sufficient conditions. For example, when the concept “Allergens” is defined by something being a protein which has a “scientific name”, a “molecular weight”, an “isoelectric point”, occurs in certain organisms and causes specific allergies, then every allergen must have the same attributes and relations and every object that has such attributes and relations must be an allergen.
- *Coherence*: the definitions given must be consistent. As Gruber notes, at the least the formally defined part of the ontology should be logically consistent. The formal part of the ontology can be checked for consistency using an inference engine provided by the knowledge representation mechanism used in the implementation stage. Gruber also adds that coherence should apply to the concepts defined informally such as those described in the ontology documentation and examples.
- *Extendibility*: the ontology should enable the addition of new concepts (as well as attributes and relations of existing concepts) without having to revise the existing conceptual definitions. This criterion affects not only the conceptualization and formalization stages but also the implementation and the maintenance stages since the design of an ontology must take into account that the ontology is a “living thing” that evolves when is used in its run-time environment. This evolution may concern not only the addition of new concepts, properties and relations but also the removal of existing ones.
- *Minimal Encoding Bias*: representational choices must not be made for the benefit of implementation. This means that an ontology must be as independent as possible from the application which will use the ontology. This will facilitate the re-use of the ontology by other applications.

- *Minimal Ontological Commitment*: define only those knowledge units¹¹ that are necessary to support the intended knowledge sharing activities. This criterion is again related to re-usability since the less committed an ontology is, the more reusable will become in new applications.

It must be stressed that although the above criteria are generally accepted in ontology design, however, like in most design problems, their application in practice will require making tradeoffs. For instance in the case of minimal ontological commitment, when the domain is narrowed it will be easier to agree with others on the ontology description. However, when the ontology provides only little knowledge for a domain this affects the range of applications that can actually use this ontology. That’s why although these design criteria drive our ontology building work, they are appropriately adapted in each separate application taking into account the requirements of the specific application. Below, we present examples of the application of the above criteria in the ontology building process.

One of our main concerns during the conceptualization stage of the ontology design is the avoidance of duplicated knowledge units. The more times a knowledge unit appears in different places in an ontology, the bigger the semantic ambiguity, thus affecting the clarity of the ontology to be built. In order to enforce the clarity criterion, we discard knowledge units that are not directly useful to the modeling of the domain in the sense that they do not provide exploitable modeling information. Figure 1 depicts a case where the knowledge units (concepts) 2 and 3 do not contribute to the modeling of the domain with exploitable knowledge therefore they are discarded (dashed lines) from the ontology. Figure 2, on the other hand, depicts another case where for clarity reasons it is necessary to further categorize the instances of the concept A by introducing the new concepts B and C. Logic-based ontology languages have the potential to derive subjective conceptualizations (define new concepts based on existing knowledge units, i.e. concept’s properties). This is extremely useful as for example a biologist could dynamically change at runtime the categorization (conceptualization) of organisms (e.g. allergens) by specifying the admissible values of their sources according to the various levels of Linnaeus’s taxonomy.

Again, during the conceptualization stage, we create a new domain sub-ontology (modularization) using a subset of domain knowledge units, whenever we consider that those knowledge units can constitute a stand-alone

¹¹We use the term knowledge units to refer to concepts, their attributes and relations.

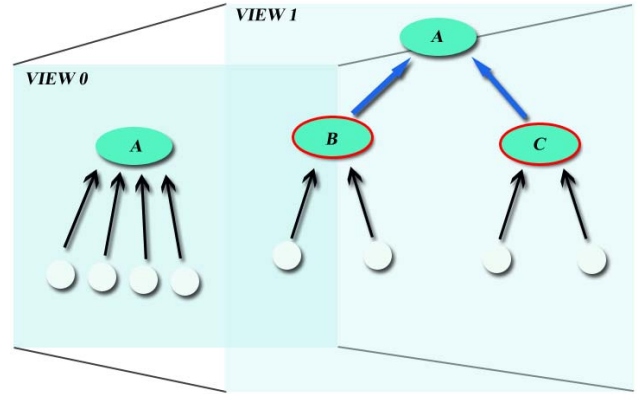
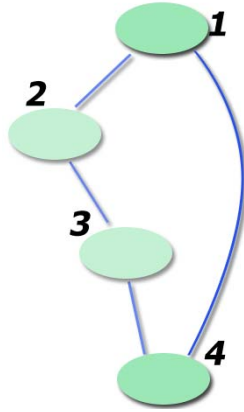


Figure 1: Redundant knowledge Units Figure 2: Defining different views

ontological model which can be associated with other ontologies. Sub-ontologies can be maintained easier than the complete ontology for the domain. They can also be shared by other ontologies in relevant applications. Such decisions are based on the minimal ontological commitment criterion. For example, the Linnaeus taxonomy can be more easily shared and maintained as a stand-alone ontology rather than being embedded in another one. This is the way we handle Linnaeus taxonomy in the ontology for the allergen domain presented in this paper.

In the formalization stage, we do not intend to “reinvent the wheel” by redefining concepts, relations and properties but based on previous work [32] on formal biomedical ontological relations, we preserve rigor and validated definitions in order to lower the semantic heterogeneity problem and speed up the ontology building process.

Concerning the implementation stage, we use the widely used description logic-based ontology language, OWL [33], a W3C recommendation since February 2004. Such languages promise more reusability, hence less cost on ontology building-ontology integration. In addition, description logic-based languages provide a knowledge representation scheme which facilitates the inspection of the ontological model for consistency, satisfiability. Furthermore they support the ontology building process by inferring implicit class hierarchy and unexpected implied relations. The domain expert should be able examine the ontology using editors like the one mentioned above throughout the design and building process and document it in order to verify its content and validate its structure and logic. Well defined and documented ontologies will most probably be adopted by other applications

and communities.

3.2 Ontology Maintenance as Ontology Population and Enrichment Task

Ontologies need to be regularly updated in order to cope with the changes occurring in the domains they model. For example, an ontology modeling the evolving domain of laptops' offers [2] must be frequently updated to include the new types of processors, screens, batteries, etc. This is also the case in biomedical domains, such as the allergen domain examined in the present work, where new allergen names or variants of existing names occur in relevant corpora and need to be included in the ontology through a maintenance process.

In this sub-section we present our methodology for the automation of ontology maintenance. This concerns mainly the population of the ontology with new instances filling their concept's attributes and its enrichment with typographic variants of existing instances. The basic idea behind this methodology is that when we want to model a domain for an application, it is enough to have at the beginning an ontology containing only a small number of knowledge units as well as a corpus of domain specific documents. It must be noted that the initial ontology is formally defined and built according to the methodology and design principles presented in the previous sub-section. This initial ontology is used to automatically annotate the corpus (*ontology-based semantic annotation*). The annotated corpus is then used to train an information extraction engine which is consequently applied on the corpus to discover new instances (*knowledge discovery*). Typographic variants of existing instances are then identified (*knowledge refinement*). Automatically acquired instances and typographic variants are presented to a domain expert who examines their correctness and adds, the correct one, in the ontology (*knowledge validation*). This is an iterative process where the new version of the ontology is employed in each iteration. The whole process is depicted in Figure 3. Each stage is presented below in more detail.

3.2.1 Ontology-based Semantic Annotation

The occurrences of ontology instances are located and annotated in the domain specific corpus using a string matching technique which is based on regular expression patterns. This technique usually exploits contextual information to disambiguate between the various interpretations an instance's attribute value may have. For example, in the allergens domain, the string

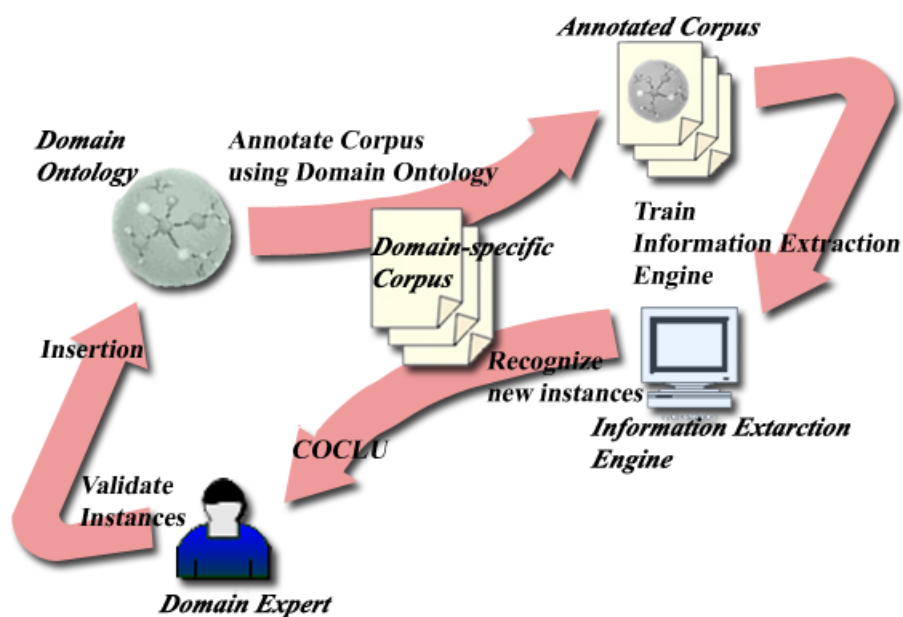


Figure 3: Ontology maintenance methodology

“9” could express an attribute value for the “Molecular Weight” or the “Isoelectric Point”. This ambiguity is resolved by the exploitation of the corresponding measurement units e.g. if “9” is followed by the string “kd” then it is a value for “Molecular Weight” and if it is followed by the string “pH” then it is a value for “Isoelectric Point”. The contextual information used for the disambiguation of such cases is already encoded into the initial domain ontology (i.e. the measurement units in case of numeric data). The outcome of this stage is a partially annotated corpus which will be used by the information extraction system at the knowledge discovery stage as training dataset. The size of the initial ontology affects corpus annotation and subsequently the training of the extraction engine. How much, is something we examine in the evaluation stage of our methodology (see next sub-section) experimenting with different sizes of the initial ontology.

3.2.2 Knowledge Discovery

The core module of this stage is a machine learning based information extraction engine which is trained over the annotated corpus. The trained engine will be able to locate then ontology instances, and numerical or time

expressions that fill instances' attributes. Various information extraction techniques can be employed in this stage. In contrast to other approaches for ontology population, our approach exploits an automatically annotated corpus in order to reduce human intervention. The new instances located by the trained information extraction engine are candidate ones since they still have to be validated by the domain expert before being included into the ontology. The size of the initial ontology, the selection of a specific extraction technique and the tuning of this technique according to the application requirements are aspects that may affect significantly the performance of the knowledge discovery stage. The testing infrastructure we use enables us to experiment with different extraction techniques as well as to easily tune the specific technique selected.

3.2.3 Knowledge Refinement

This stage aims to enrich the ontology with a non-taxonomic lexico-semantic relation holding between concepts instances and their typographic variants that may occur in the domain specific corpus. The acquisition of this relation is based on the intrinsic characteristic of the instance, i.e. the string that represents it in the corpus. We name this relation “*hasTypographicVariant*” and it seems very promising in the conceptual modeling of biomedical domains since the names of many entities are written in a non-standardized name convention nomenclature [34]. For example, Hanisch et al [35] notice this in the gene domain. The identification of different typographic variants of existing instances is performed by a *novel compression-based clustering algorithm*, named *COCLU (COmpression-based CLUstering)* [36]. The algorithm is based on the assumption that different lexicalizations of an instance (typographic variants) use a common set of ‘core’ characters. Therefore, typographic variants that are ‘close’ to this set are potential alternative expressions of the same instance, while those ones that are ‘far’ from this set are potentially related to a different instance. COCLU is a partition-based clustering algorithm which divides the data into several clusters and searches the space of possible clusters using a greedy heuristic. Each cluster is represented by a model, rather than by the collection of data assigned to it. The cluster model is constructed incrementally, since the algorithm dynamically generates and updates the clusters by processing a string at a time. COCLU employs an innovative score function that measures the compactness and homogeneity of a cluster. This score function is termed Cluster Code Difference (CCDiff) and is defined as the difference of the summed length of the coded string tokens that are already members of the cluster, and the

length of the same cluster including the candidate string. A string belongs to a particular cluster when its CCDiff is below a specific threshold and it is the smallest one among all the CCDiff's of the given string with all existing clusters. A new cluster is created if the candidate string cannot be assigned to any of the existing clusters. As a result, it is possible to use the algorithm even when no initial clusters are available. Similarly to many incremental algorithms, the order in which the strings are encountered influences the performance of the proposed algorithm. For this reason, COCLU iteratively computes the CCDiff for all the existing clusters and for all candidate strings and selects the one that can be more reliably assigned to a cluster.

3.2.4 Knowledge Validation

At this stage, the domain expert validates the extracted instances' attribute values and typographic variants that have been derived from the knowledge discovery and refinement stages. The validated information is then added to the ontology. The outcome is a new version of the ontology updated with knowledge extracted from the domain specific corpus. A new iteration begins with the new version of the ontology. The iterative process will stop when no more changes in the ontology are possible.

3.3 Evaluation

Concerning the evaluation of the ontology design, in terms of the design criteria applied, this is mainly performed through the formal model and the knowledge representation mechanism employed. This sub-section presents our methodology for evaluating ontology maintenance which is crucial for the successful tuning of the modules employed. The first step in the evaluation methodology is the collection of the domain specific corpus which must be performed carefully according to certain pre-specified criteria. The collected corpus is then semantically annotated by domain experts using the complete version of the ontology. It must be noted here that in order to evaluate the ontology maintenance process, we must be able to compare the various versions of the ontology (produced at the end of each iteration) against a *gold ontology* which encodes the domain knowledge of the whole domain specific corpus. This *gold ontology* is built by domain experts and is employed in the corpus annotation process. The annotation is done using a well-established methodology which is supported by domain-specific guidelines and user friendly annotation tools to assure the consistent annotation of the corpus. The final step of the evaluation methodology involves the setting

of various experimental parameters. For instance, as it was noted in the previous section, we need to study the effect of the size of the initial ontology. This can be achieved by setting up experiments with initial ontologies of different sizes. Each of the above steps is presented below in more detail.

3.3.1 Corpus Collection

The basic idea behind the evaluation methodology is to measure the performance of our maintenance approach in terms of the knowledge acquired when it is applied in a domain specific corpus compared to the knowledge encoded in an initial ontology of limited size. This initial ontology must represent the domain knowledge at a specific point in time. This is necessary to be done in order to simulate, in our experiments, the maintenance process in a real-life scenario. For instance, in the case of allergens the starting point might be an ontology representing the domain knowledge from a corpus of PubMed abstracts published by the end of 2000. The aim of the simulation experiments in such a case would be to populate and enrich the initial ontology with knowledge acquired from PubMed abstracts published by the end of 2002. The setting of such experiment requires the collection of PubMed abstracts for allergens published by the end of 2002, the ordering of these abstracts according to their publication dates and the creation of a corpus subset containing those ones published by the end of 2000. It also requires the manual building of a formally defined ontology encoding all the knowledge acquired from the complete set of abstracts (gold ontology) and a subset of this ontology encoding the knowledge from the subset of abstracts.

Therefore, corpus collection and organization is crucial for the experimental setting and it must be done carefully. The corpus is studied and some statistical figures are collected in terms of the amount of occurring instances. A representative corpus must contain adequate amounts of all the types of instances and attribute values. Otherwise during knowledge discovery the trained information extraction engine won't be able to locate accurately information that is not represented adequately in the training corpus.

3.3.2 Corpus Annotation

The purpose of a corpus annotation methodology is to establish common definitions of the human annotation task. Based on previous work on corpus annotation in the CROSSMARC¹² R&D project [37] and the lessons learned

¹²<http://www.iit.demokritos.gr/skel/crossmarc>

from the annotation of the allergen corpus we specified the major steps and principles of an annotation methodology which can be applied in corpora from similar domains and tasks.

The corpus annotation methodology that has been developed and followed in this project is comparable to standard annotation practice for the purposes of Information Extraction [38, 39]. The annotation task is based on annotation guidelines that are issued for a specific domain in order to ensure a common understanding between annotators of what is to be annotated and how it should be annotated. Domain experts decide on the important features that characterize a domain specific entity (based on the ontology for the specific domain). The human annotators proceed with the annotation of these features in the corpus following the annotation guidelines. These guidelines name the conventions that must be followed in the annotation of the entity features in order to ensure consistency in the annotation of a corpus. Feature annotation guidelines contain suggestions for the annotation of all types of features when the desired annotation is not obvious, as well as suggestions for the annotation of specific features with examples. No matter how concise or inclusive the annotation guidelines are, there is always the chance that some annotations may deviate from them for various reasons, or some other annotations may be missing due to human errors. In order to minimize such errors there are two human annotators involved and their annotations are compared. The differences observed are used as clues for the creation of the final annotations. These clues can be useful for spotting cases that have not been mentioned in the guidelines, instructions that confused the annotators, cases where the guidelines have not been followed and random mistakes. The results of the comparison are incorporated in a report that gives indications for the creation of the final annotations.

3.3.3 Setting the parameters of the experiment

Following the collection of the corpus and the ontology-based annotation, we still have to set the following parameters of each experiment:

- Size of the initial ontology: this is determined by the chronological period we want to cover, that is by the subset of the corpus which contains documents up to a specific point in time. We can create initial ontologies corresponding to different time periods in order to examine their effect to the final results. Two such experiments were performed in the case of the allergens ontology (see section 7).

- Information extraction technique: there are various machine learning based techniques that can be employed in the knowledge discovery stage. Selection of such a technique depends on the characteristics of the technique and the specific domain. The tuning of these characteristics affect the performance of the overall method. The technique we use in the present work is first-order discrete Hidden Markov models (see section 5.2).
- Tuning of the clustering algorithm COCLU: the application of COCLU depends on the characteristics of the domain. For example, there may be types of instances or attribute values without typographic variants at all or with only a few occurrences of them. In addition, since the order with which the strings are fed to COCLU may affect significantly its performance, this must also be examined, employing the functionalities provided by COCLU for this purpose.

4 Building the Allergens Ontology

This section presents the decisions taken during the building of the allergens ontology following the design principles presented in section 3.1. It also discusses the problems encountered and the solutions given. Then it presents the allergens ontology along with its main characteristics.

4.1 Ontology Design

We present our experiences on modeling the allergens domain exploiting existing ontologies (semi-formal models) and databases, as well as documents that describe the allergen nomenclature system. We believe that these experiences can be helpful also for other biomedical domains which currently have similar semi-formal models. Moving from a semi-formal model to a formal one for a domain ontology is a necessary step towards the efficient exploitation of the domain specific knowledge.

To implement the allergens ontology we used the description logic-based ontology language OWL. Specifically, we adapted the OWL DL sublanguage of OWL in order to take advantage of its consistency and satisfiability checking mechanisms, and its reasoning and inference capabilities. We also used the OWL Protégé’s plug-in [40] to support ontology editing in OWL. Figure 4 (taken from IUIS allergen list) depicts allergens along with some characteristics i.e. their name, biomedical id or obsolete name, species name where

Species name	Allergen name	Biochemical id or obsolete name	MW kDa SDS- PAGE	C: cDNA P: peptide sequence	Reference and/or accession number
A. Weeds					
Asterales					
<i>Ambrosia artemisiifolia</i>					
short ragweed	Amb a 1	antigen E	38	C	8, 20
	Amb a 2	antigen K	38	C	8, 21
	Amb a 3	Ra3	11	C	22
	Amb a 5	Ra5	5	C	11, 23
	Amb a 6	Ra6	10	C	24, 25
	Amb a 7	Ra7	12	P	26
<i>Ambrosia trifida</i>					
giant ragweed	Amb t 5	Ra5G	4.4	C	9, 10, 27
<i>Artemisia vulgaris</i>					
mugwort	Art v 1		27-29	C	28
	Art v 2		35	P	28A
	Art v 3	lipid transfer protein	12	P	53
	Art v 4	profilin	14	C	29
<i>Helianthus annuus</i>					
sunflower	Hel a 1		34		29A

Figure 4: Segment of the list of allergens as of September 01, 2004. (WHO/IUIS)

they can be found, molecular weight etc. A fundamental question in ontology building is which are the relevant and irrelevant terms that constitute the concepts of a domain ontology. We acquire from the allergen list only information that is explicitly related to the modeling of the allergens domain in order to conform with the minimal ontological commitment design criterion and the non subjective modeling. Thus, the field “reference and/or accession number” is not used. The names of the top-most classes used in this classification are: weeds, grasses, animals, fungi, insects, foods, trees and others. This is a subjective classification of genus which increases the semantic ambiguity as several species can be classified under these classes e.g. the species “*Betula verrucosa*” (its common name is birch) can be classified either as tree or food. Thus, these classes are not included in our ontology.

We noticed that some branches of the taxonomy appearing in the IUIS allergen list adopt a small part of the Linnaeus taxonomy, specifically the levels order, genus and species. Following the extendibility and minimal ontological commitment design criteria, we created separately the Linnaeus

ontology and imported it into the allergens ontology linking the source of allergen (species) with the corresponding species at the Linnaeus ontology. The Linnaeus taxonomy provides labels for the groups denoting the taxonomical levels (e.g. Domain, kingdom etc.). Classes (e.g. Eukarya, Plantae, Animalia, Homo) appearing in each level are used to classify the various organisms. We use the labels of Linnaeus taxonomy as concepts' labels in our Linnaeus ontology and we treat their corresponding classes as instances to the concepts. We adopt such a design view to solve three practical problems during the integration of Linnaeus ontology with the allergens ontology. Firstly, our ontology is compatible with OWL DL since it accepts instances as property values. Secondly, beyond using species instances as allergen's source values, genus instances (e.g. Homo) can also be used. Thirdly, the ontology can be easily and dynamically updated with new instances (new biology classification's classes). Figure 5 illustrates the Linnaeus ontology. A fundamental issue in ontology design, which is related to the clarity and

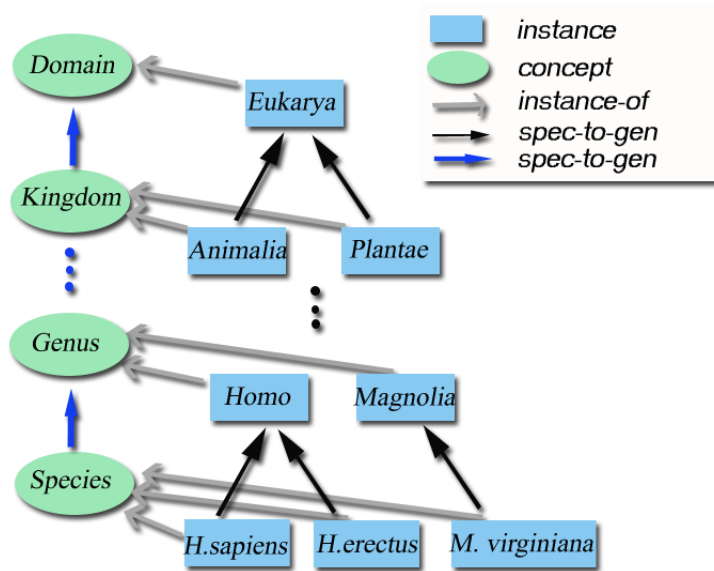


Figure 5: The Linnaeus Ontology

minimal ontological commitment design principles, is the distinction between primitive and defined concepts. Primitive concepts are axiomatically accepted, whereas defined concepts are based on the primitive ones. The correct ascription of a property to the right concept is crucial for the definition of a defined concept and results in a well-designed domain model.

4.2 The Allergens Ontology

A biochemist and a knowledge engineer were involved in the creation of the allergens ontology following the design principles stated above and acquiring knowledge from the IUIS allergen list and the allergen nomenclature. The ontology was implemented in the OWL representation language using the Protégé OWL plug-in [40]. The ontology will become soon publicly available.

Figure 6 illustrates the allergens ontology. Ellipsis stands for concepts whereas arrows denote ontological relations. Gray ellipsis denotes a defined concept. The “*Allergens*” concept is subsumed by the “*Proteins*” concept. Moreover, “*Allergens*” are linked to “*Proteins*” through the relation “*hasBiochemicalIdentityAs*” which is the inverse of the “*hasAllergicActivityAs*” relation attached to “*Proteins*” in order to associate instances between those two concepts. We further divide “*Allergens*” in two defined concepts using the “*is-a*” relation, driven mostly by the maintenance task and the corpus knowledge. In some abstracts an allergen is named according to the allergen nomenclature and in some others an allergen is described providing some of its characteristics i.e. protein, molecular weight, isoelectric point, source. We characterize as “*Named Allergens*” those that have scientific name derived from the allergen nomenclature and as “*Descriptive Allergens*” those that do not have scientific name. The “*Named Allergens*” subsume “*Iso Allergens*” which further subsume “*Variants*” These concepts are defined based on the restrictions imposed by the existing relation “*hasAminoAcidSimilarityWith*” between allergen instances and the allergen concept e.g. “*Iso Allergens*” “*is-a*” “*Allergens*” that have “*hasAminoAcidSimilarityWith*” above 67% with other allergens instances. This similarity relation is subsumed by the transitive relation “*hasSimilarityWith*” which is being restricted in the “*Iso Allergens*” and the “*Variants*” concepts. “*Allergen Sources*” instances are connected through the “*occursIn*” relation with “*Allergens*” instances. Moreover, the “*Allergen Sources*” are connected with the appropriate level of the Linnaeus ontology and the instances of “*Allergies*” type through the “*causes*” relation. The concept “*Allergens*” has also the “*hasVariant*” and “*hasIsoAllergen*” relations which connect allergen’s instances with instances from the “*Iso Allergens*” and “*Variants*” concepts.

The attributes of the allergens ontology’s concepts are centralized in table 1. The “*Allergens*” concept has the attributes “*Scientific Name*” which holds the name derived from the allergen nomenclature, the “*Common Name*”, which hold the commonly used name, the “*Former/obsolete Name*”, which holds the name we used to have before the setting of the allergen

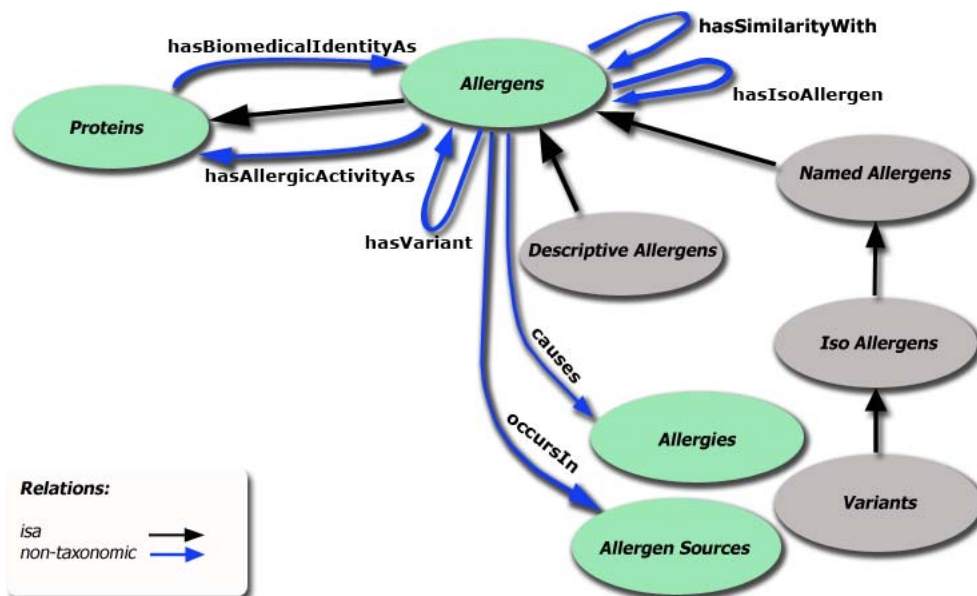


Figure 6: Allergens ontology: concepts and relations

nomenclature system, the “*Molecular Weight*” and the “*Isoelectric Points*”, which hold their numeric values. The “*Proteins*” concept has the attribute “*Name*” and the “*Allergen Sources*” concept has the attributes “*Scientific Name*” and “*Common Name*”. Finally, the “*Allergies*” concept has the “*Name*” attribute in which the allergy’s name is assigned.

5 Ontology Maintenance

New allergen names appear constantly in the literature creating the need for maintaining allergen specific resources. We applied our methodology for ontology maintenance in the allergen domain in order to populate the allergens ontology with new instances located in relevant PubMed abstracts and enrich it with typographic variants of the instances located in these abstracts. The application of each stage of the methodology is presented in the following sub-sections.

5.1 Ontology-based Semantic Annotation

The occurrences of ontology instances in the domain specific corpus (PubMed abstracts) are located and annotated using a string matching technique

Table 1: Allergens ontology: concepts and their attributes

Concepts	Attributes
Allergens	<i>Scientific Name</i>
	<i>Common Name</i>
	<i>Former/obsolete Name</i>
	<i>Molecular Weight</i>
	<i>Isoelectric Point</i>
Proteins	<i>Name</i>
Allergen Sources	<i>Scientific Name</i>
	<i>Common Name</i>
Allergies	<i>Name</i>

which is based on regular expression patterns. For each instance and its typographic variants a case-insensitive pattern is built. For example the ontology instance “Cor a I” is converted into “\sCor\sI\s”.

A rule followed in the application of the patterns is to select the maximum spanning annotated lexical expressions. This rule handles problems created by overlapping instances in the ontology. For example, the text segment “Cor a 1.01” will be annotated correctly by the instance “Cor a 1.01” but erroneously by the instance “Cor a 1”. This is also the case for the lexical values of the concept attributes. However, in case of numeric values we match the numeric lexical expression and we use disambiguation techniques to decide on the correct annotation, i.e. “Molecular weight” vs “Isoelectric Point”. Context-typed information is used to disambiguate between the various roles a numeric value can have. For example, the numerical expression “5” could be a value for the “Molecular Weight” or the “Isoelectric Point” attribute. Those ambiguities are resolved by the exploitation of the measurement units e.g. if “5” is followed by the string “kd” then it is a value for “Molecular Weight” and if is followed by the string “pH” then it is a value for “Isoelectric Point”. Also, the valid numeric ranges of the attribute values are exploited in the disambiguation process. For example, the isoelectric point cannot exceed the value “14”, hence the string “32” cannot be a value for the “Isoelectric Point”. We should note that all the information involved in the disambiguation process is encoded in the domain ontology.

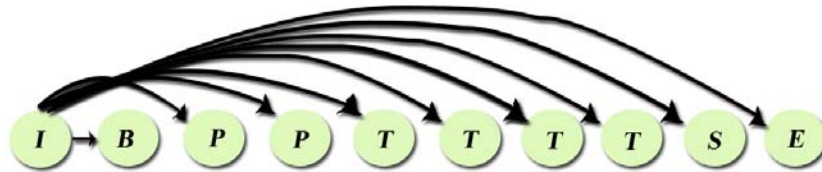


Figure 7: Sample of HMM’s structure

5.2 Knowledge Discovery

This stage employs a machine learning based information extraction engine which is trained over the annotated corpus. The trained engine is then used to locate new ontology instances. The extraction techniques used in this work is first-order discrete Hidden Markov Models (HMMs). A single HMM is trained for each attribute type (see Table 1) as proposed in [41] and [42]. HMMs exploit tokens to learn the context in which the values of a particular attribute occur in. The structure of an HMM consists of six different type nodes. These nodes are associated with the HMM’s states through one-to-one mapping. The *start (I) node* and the *end (E) node* model the first token and the last token of the abstract, respectively. The *target (T) node* models the tokens that represent the instance. The *prefix (P) node* models the tokens that appear directly before the tokens that represent the instance whereas the *suffix (S) node* models the tokens following the tokens of the instance. Finally, the *background (B) node* models all the other tokens. Except form the start, end and background types of nodes, the number of other types of nodes is set by hand. For example, the HMM structure depicted in Figure 7, involves two prefix nodes (i.e. we examine the two tokens on the left of the candidate instance), four target type nodes (i.e. a candidate instance may consist of four tokens at most) and one suffix type node (i.e. we examine one token following the candidate instance). The arrows represent the transitions between the nodes. For simplicity, Figure 7 depicts only transitions from the “I” node to all other nodes, whereas there are transitions between other nodes as well. The structure of each HMM varies since we chose the one that performs best after a quick experimentation.

The transitive probabilities are estimated in a single pass over the training dataset by calculating ratios of counts (maximum likelihood estimation). At runtime, each HMM is applied to one document in the corpus, using the Viterbi procedure to identify matches. This ontology-driven machine learning approach differs from the classical supervised methods as it does not use human-provided training examples but examples provided by the domain

ontology.

5.3 Knowledge Refinement

This stage aims to enrich the ontology with a non-taxonomic lexico-semantic relation holding between concepts instances and their typographic variants that may occur in the domain specific corpus (*“hasTypographicVariant”* relation - see section 3.2.3). This relation links term variations caused by typographical phenomena (“Cor a I” and “Cor-a-I”), morphological phenomena (e.g. “P Centrinum” and “P. Centrinum”) or lexical phenomena (e.g. “Amb a 1” and “Amb alpha 1”) Also, acronyms/abbreviations (e.g. “Small Rubber Particle Protein” and “SRPP”) cases are covered by this relation.

In order to enrich the ontology with this relation, we use the clustering algorithm COCLU presented in section 3.2.3. COCLU’s use is two-fold. Firstly, it can be used as a classifier which assigns an ontology instance lexicalization to the appropriate group. In this way COCLU is being used for discovering typographic variants of existing ontology instances. In addition, it can be used for discovering new groups, beyond those denoting the already known instances. A group is defined by an ontology instance and its typographic variants. In table 2, we see potential ontology instances and their typographic variants.

Table 2: Examples of ontology instances and their typographic variants

Instance Name	Typographic variant 1	Typographic variant 2
Amb a I	Amb alpha I	Amb a 1
Cor a 1	Cor a I	Cor-a-I
Fel d I	Felis domesticus I	Fel d1
Penicillium citrinum	P citrinum	P. citrinum
small rubber particle protein	SRPP	-

5.4 Knowledge Validation

In this stage the domain expert validates the extracted instances and typographic variants using a visualisation tool of the Ellogon text engineering platform [43]. The expert inserts then the validated information into the ontology using the Protégé ontology editor. We help the domain expert to validate information by highlighting relevant information on the abstract. At the end of this stage a new iteration begins with the updated ontology. Iterations stop when no more changes in the ontology are possible.

6 Experimental Setting and Results

We evaluated the performance of the incremental ontology maintenance methodology in terms of its capability to discover new ontology instances and enrich the ontology with the lexico-semantic relation “*hasTypographicVariant*”. We would like to stress that we do not evaluate the performance of the information extraction system to discover occurrences of instances in abstracts but we evaluate the discovery of instances and their typographic variants against a gold ontology that has been manually constructed from the corpus used.

6.1 Corpus Collection and the Gold Ontology

The corpus was collected from PubMed abstracts posing a query with keywords related to the thematic domain at hand (allergens). The collected corpus was then examined by domain experts to verify the content of each abstract against the allergen domain. This process resulted in 279 abstracts that describe allergens. The dates of abstracts span from 1974 to 2003.

Table 3 provides information concerning the occurrences of instances (annotations) in the corpus. Although the corpus has been methodically annotated (see section 6.2) at a fine grain using the domain specific ontology, for the specific task we only exploit the annotation top-most types. From this

Table 3: Statistical information of the corpus

Allergens Scientific Name	Proteins Name	Allergies Name	Allergen Sources Scientific Name	Total
1645	856	183	1973	4657

domain ontology a gold ontology was derived keeping only those instances that appear in the corpus as well as their typographic variants. In table 4 we give statistical data concerning the unique instances of the gold ontology as well as their typographic variants.

Since we developed our gold ontology from instances that occur in the selected abstracts (corpus-based ontology building), the ontology does not uniformly contain the formal names of each instance. For instance, the ontology could contain as allergen source name the typographic variant “*P. notatum*” instead of its formal name “*Penicillium notatum*” which does not appear in the corpus. Hand crafted ontologies mainly contain formal names

Table 4: Information concerning the gold ontology

	Allergens Scientific Name	Proteins Name	Allergies Name	Allergen Sources Scientific Name
Number of Unique Instances	252	151	106	40
Number of Typographic Variants	59	43	79	14

of entities. If such ontologies are used as input to our incremental ontology maintenance approach, they can be enriched with typographic variants, thus resulting in more realistic, corpus-based ontologies.

6.2 Annotating the Allergens Corpus

The 2nd step in our evaluation methodology (see section 3.3) is the semantic annotation of the collected corpus by domain experts using the gold ontology. This subsection presents the annotation tool, the domain specific guidelines and the labels used in the annotation of the allergen corpus.

6.2.1 The Annotation Tool

The annotation of the corpus was accomplished using the Ellogon language engineering platform [14] and more specifically its tool for adding and editing annotations (Fig. 8). All instances are annotated as many times as they appear in the abstract, beginning from the abstract title. The annotator marks the text segment (string) that wants to annotate and ascribes an annotation type to it by clicking on the corresponding button on the right hand side of the annotation tool. The annotation buttons have been organized in order to ease and accelerate the annotation process. We have categorized the annotation types into 5 categories. Two of them refer to the scientific and common name of the allergen’s source. Two more discriminate the annotation types that refer to the attributes of named and descriptive allergens and the last one groups annotation types that refer to common allergen’s attribute along with the annotation types for the protein’s and allergy’s attributes. Fig. 8 depicts the annotations that exist in the abstract by uniquely highlighting the various text segments. For example the string “guinea pig” is annotated with the allergen’s source common name

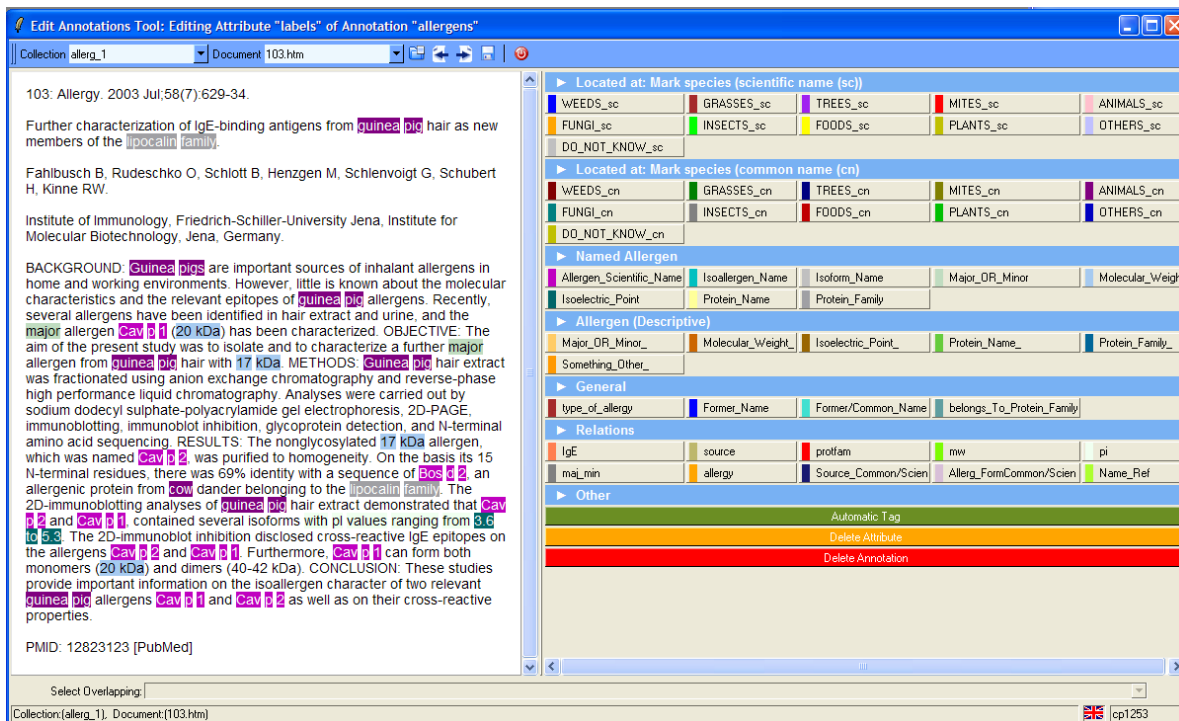


Figure 8: Tool for editing annotation

annotation type and the string “20 kDa” is annotated with the annotation type “Molecular Weight”. Although we provide annotations (see buttons in figure 8) which further categorize the possible allergen’s source, we do not take these subcategories into account at the maintenance phase as it is a rather subjective categorization of the “Allergen Sources”.

6.2.2 Types of Annotations

The annotation labels, i.e. types of annotations, come from the labels of the ontological components (see Table 1 presenting the concepts and their attributes). Precisely, these domain specific labels are:

- the *scientific name* (Genus and species according to the Linneaus system of nomenclature) and *common name* of the allergen’s source,
- the allergen’s *scientific* and *common/former names*,
- its *molecular weight* and *isoelectric point*,

- the *protein family* to which the allergen possibly belonged or showed great homology (indicative of its biochemical function/identity) or even the protein name that happened to be an allergen and, finally,
- the *type of allergy* it elicited.

As many allergens do not yet have a scientific name, we divided allergens into named and descriptive ones providing two different sets of annotation labels. Both have common annotation labels except from the scientific names label which concerns only the named allergens.

6.2.3 The Annotation Guidelines-General Guidelines and Problems

As far as the *scientific name of the source* is concerned, any instances such as *Dermatophagoides pteronyssinus*, *Olea europea*, *Plantago lanceolata*, *Blatella germanica* or *D. pteronyssinus* were annotated; in cases where the abstract referred to the species of a whole genus (i.e. *Aspergillus* spp.) we decided to exclude it from the annotation. Attention should be paid not to annotate bacterial names such as *E. coli* as sources' names, for they indicate the biological systems in which the allergens were expressed in order to be studied. The annotation of the sources' common name was not very difficult or complicated.

As already mentioned, we divided allergens into *named* and *descriptive*. If the scientific name of the allergen was mentioned in an abstract, then all the other instances that represented properties of this allergen (molecular weight, isoelectric point, protein name, protein family, major or minor) were annotated in the category "Named Allergens". In case the allergen's scientific name was not given in the abstract, but the allergen was "described" by a known protein name-for example profilin-or an unrecognizable name such as ABA-1, Ra-5, AgDg1, then our first priority was to search for this name on the aforementioned list of allergens. If this name was found on the list, we considered this allergen as a named allergen, since it has been given a scientific name but simply the authors of the abstract did not mention it. We continued by annotating all the instances defining this allergen in the "Named Allergens" category. In case this name was not found on the list, we considered this allergen as a descriptive allergen, which has not yet been scientifically named, therefore we annotated all the instances describing it in the "Descriptive Allergens" category. All other proteins mentioned in the abstract as having homology to the allergen were annotated only if they were allergens themselves (found on the list, if not mentioned in the abstract).

As far as the *allergen's scientific name* is concerned, we annotated every name such as “Amb a 1” or “Amb a I”, where the first three letters represented the first three letters of the source’s genus (*Ambrosia*), the letter separated by spaces was the first letter of the source’s species (*artemisiifolia*) and the Arabic or Roman numeral indicated the allergen in the chronological order of purification (1, I, first). Despite the fact that the Roman numeral has been replaced by an Arabic one in the official Allergen Nomenclature, most authors/scientists have not yet complied with this change and insist on referring to allergens with their obsolete “scientific” names. In most cases this is not a problem. However, there are a few cases where the Roman numeral has been replaced by an Arabic one that does not have same value and therefore does not indicate the same order of identification (i.e. Lol p 5 was previously known as Lol p IX). In most cases we annotated names containing both Roman and Arabic numerals as scientific names and only in the cases where there has been a change in the numeral’s meaning we annotated the name with the Roman numeral as “Former/obsolete Name”. Another problem - yet an extreme case - we confronted was in cases such as “...Sol i ...I, II, III and IV)...” where the first two parts of the allergen’s scientific name were separated from the numeral by two or three words and therefore we did not annotate such cases.

We annotated as “*Former/obsolete Name*” all the names referring to an allergen that were used by scientists before the allergen was given a scientific name; those names are often acronyms containing a part of the source’s name (i.e. Ra 6, Ra 3) or other names including information on the allergen’s biochemical identity (i.e. antigen E, now known as Amb a 1). Many scientists are still using these names for the description of an allergen due to the fact that they are more familiar with these terms rather than the official scientific names.

As far as the *molecular weight* is concerned, we annotated a number or a range (i.e. 14-20) - implying that the allergen represents a group of highly homologous isoallergens and isoforms - or even a whole expression (i.e. between 42 and 48) accompanied by the measurement unit (Kd, kD, Kda, kDa, Daltons, kilodaltons etc). The problems we had to cope with were mainly two. The first one was due to the fact that in many cases the authors referred to the molecular weight of the allergen as a monomer (molecular weight derived from SDS-PAGE) or a dimer (molecular weight calculated from HPLC). We circumvented this by annotating the molecular weight referring to the monomer, but unfortunately lost the information on whether the allergen usually appeared as monomer or dimer (or other polymers). A second point at which we had to pay attention was the char-

acterization of a molecular weight as “apparent” or “determined” and the fact that the authors frequently referred to the molecular weight of the “mature” protein. In this situation we did not annotate the apparent molecular weight. In some cases we found within an abstract molecular weight referring to a fragment of an allergen (N-terminal, C-terminal fragment) or to an incomplete gene product of a recombinant allergen; such molecular weights were not annotated.

There were no significant problems concerning the annotation of the *isoelectric point* and the *type of allergy*. We could only mention that in many cases the abstracts referred to a “pH level” and not clearly to the isoelectric point (pI) and that in few cases an allergen was characterized as “main” or of “medium importance” or also as “intermediate”. The “type of allergy” refers to medical symptoms such as asthma, rhinitis and atopic dermatitis or the names of allergic syndromes, for example oral allergy, latex-fruit allergy and pollinosis. We did not annotate anything containing information on the types of allergic reactions elicited - such as type I hypersensitivity - as they are types of immune responses.

Concerning the characterization of an allergen as major or minor, it is important to note that in most cases of our corpus it was clearly mentioned in an abstract; there were few cases where something like this was not mentioned and we therefore did not annotate any other phrases that could give us such information (for example percentage of IgE binding).

As most allergens are proteins, we decided to have two features containing information on their biochemical function/identity: *protein name* and *protein family*. The major problem we faced was the annotation of the protein name or protein family the allergen belonged to, as there was a conflict between the two fields. A real problem, not only for us but also for most biochemists and biologists, is the fact that protein names show considerable variation because of the existence of multiple naming conventions (based on function, sequence features, gene name, cellular location, etc.). Therefore, in many cases we could not discern a protein name from a protein family (i.e. profilin and family of profilins). The annotation at that point was based on the general context of the abstract; for example, we annotated profilin as a protein name at those points where we understood that the authors were referring to a specific profilin (i.e. cherry profilin) and as a protein family at those points where the authors were referring to profilin as a general attribute (as a member of a whole), indicating the biochemical function/identity of the allergen. In general, we annotated as protein name and protein family anything that was related to function. For example glycoproteins or characteristics concerning structure were not annotated.

6.3 Setting the Parameters

We evaluate the performance of the proposed ontology maintenance method simulating the evolution of the allergens ontology by the creation of two initial ontologies that corresponds to different time periods. Thus, we divided the corpus chronologically based on abstracts' dates and we constructed two initial ontologies which reflect the knowledge contained in those portions of the corpus. The two ontologies were constructed from the 25% and 50% of the total number of abstracts that had been chronologically sorted. The number of their initial concept instances used in the experiment is illustrated in table 3. The first ontology contains the 16.72% of the total number of "Allergens Scientific Names" in the gold ontology whereas the second initial ontology contains the 42.76%. In general, the first ontology contains the 15.59% of the total number of gold ontology's instances and the second one contains the 34.94%. These two ontologies were used as different starting points to our maintenance method in order to simulate an ad hoc application of the proposed methodology.

It is worth noting that we do not work on the discovery of instances of the attributes "Isoelectric Point", "Molecular Weight" and "Former/obsolete Name" as the set of their values are closed and predefined in the ontology. The performance of the maintenance method passes through the appropriate selection of the algorithms used in the various stages of the methodology as well as their parameters' tuning. Specifically, we have selected to use HMMs as information extraction engine and an unsupervised machine learning algorithm (COCLU) for discovering typographic variants of instances. A good selection of these parameters affects the performance of the ontology maintenance process and consists firstly in the structure of the HMM (see section 5.2) and secondly on the correct grouping of instance (see section 5.3). We spend some time experimenting with some parameter values and we chose the one that gave the best results.

6.4 Evaluation of the Ontology Maintenance Task

6.4.1 Ontology Population Evaluation

In this experiment, we evaluated the improvement of the results at each iteration simulating the evolution of the domain. As noted before, we conducted experiments using two different initial ontologies (see section 6.3) to measure the effect of the initial ontology size in the maintenance process. In each experiment, the initial ontology was used to annotate semantically the corpus. The annotated corpus was then used to train the extraction engine.

Each trained engine was then applied to the whole corpus. As it is shown in Table 5, in the case of the 1st ontology the extraction engine managed to locate 168 more “Allergens Scientific Name” instances, whereas in the case of the 2nd ontology it managed to locate all the “Allergens Scientific Name” instances from its first application.

Table 5: Ontology Population Evaluation

		Allergens Scientific Name	Allergen Sources Name	Proteins Name	Allergies Name	Total
1st ontology	% of the Gold Ontology	16.72	19.1	5.4	31.4	15.59
	Initial Instances	52	37	10	17	116
	Target Instances	311	194	185	54	744
	0th Iteration	168	23	43	7	241
	1st Iteration	74	31	8	4	117
	2nd Iteration	12	13	8	1	34
	% final Coverage	98.3	53.6	37.3	53.7	68.3
2nd ontology	% of the Gold Ontology	42.76	36.60	18.92	38.88	34.94
	Initial Instances	133	71	35	21	260
	Target Instances	311	194	185	54	744
	0th Iteration	178	73	55	12	318
	1st Iteration	-	11	11	3	25
	2nd Iteration	-	6	1	2	9
	% final Coverage	100	83.0	55.1	70.3	81.9

After this first application (0th iteration column), the updated ontology is used in the new iteration. The results (1st iteration column) show that more instances were found leading to a new update of the ontology which is then used in another iteration (2nd iteration column). At the end of the 2nd iteration the updated ontology derived from the 1st ontology covers almost all the “Allergens Scientific Name” instances and some of “Allergen

Sources Name”, “Proteins Name” and “Allergies Name” instances (53.6%, 37.3% and 53.7 respectively). The results are even better in the case of the updated ontology derived from the 2nd ontology. Starting with only 34.94% of the gold ontology’s total number of instances, the method succeeds to populate the ontology increasing its coverage to 81.9% in 2 iterations. The system performs better and faster in the 2nd ontology case since the more instances in the initial ontology, the more types of contexts are found and more training examples are provided to the extraction engine.

The results are not so good for the “Proteins Name” and “Allergies Name” cases. In “Proteins Name” this is due to the small number of training examples (5.4% in the 1st experiment), whereas in “Allergies Name” this is due to the small number of instances (54 in total). In the experiment with the 2nd initial ontology, the number of training examples for “Proteins Name” increases but not as much as the other types of instances. So, the problem of limited training examples remains. In “Allergies Name”, on the other hand, only 4 new instances are added (from 17 to 21), so the situation remains practically the same.

It is worth noting that a study on the evolution rate of a domain could indicate the exact time period in which the proposed methodology should be applied for keeping up-to-date the ontological knowledge and having the highest performance. Special attention should be given to the ontology-based annotation of the corpus since the annotated corpus will be used to train the information extraction engine. Restricting the task to a concrete domain and using domain-independent techniques we correctly managed to find the lexicalization of ontology instances in the corpus and to provide iteratively a valid training corpus.

6.4.2 Ontology Enrichment Evaluation

This experiment measures the performance of the COCLU algorithm, which is used to group typographic variants of existing ontology instances. This algorithm is useful for reducing domain expert’s workload in validating candidate instances and typographic variants. We conducted this experiment by asking COCLU to categorize the annotated instances to classes defined by an ontological instance and its potential typographic variants. COCLU assigns typographic variants to corresponding classes whereas new classes are created when there is not sufficient evidence for similarity between an annotated instance and the initial classes. We configure this similarity with COCLU’s threshold value.

Table 6 presents the results from the application of COCLU in the whole

corpus. The classes columns denote the number of the classes we are trying to identify (target classes), the classes produced by COCLU in total and the correct classes within them. The target classes are defined by the number of the unique instances in the gold ontology (see Table 4). The variants columns denote the number of the target typographic variants (see Table 6) and the number of variants correctly produced by the algorithm.

Table 6: Ontology enrichment evaluation results

	Target Classes	Classes Produced	Correct Classes	Target Variants	Correct Variants
Allergens Scientific Name	252	252	252	59	5
Allergen Sources Name	151	121	121	43	34
Proteins Name	106	125	106	79	19
Allergies Name	40	40	40	14	10

COCLU performs quite well in the identification of classes. This is mainly due to the setting of a low threshold value (different for each type of instance) which enables the creation of a separate class for each instance even in the cases where the instance differs only in one character from instances belonging in other classes.

However, the setting of a low threshold value affects the identification of variants, especially in those cases where there are certain domain specific peculiarities. The algorithm manages to locate most of the “Allergen Sources Name” variants (34 out of 43), producing only correct classes. It also performs well in “Allergies Name” managing to find all the target classes and almost all the variants (10 out of 14). But in the case of “Allergens Scientific Name”, the algorithm has a very poor performance (5 out of 59). This is mainly due to the fact that it was confused by the isoallergens of several allergens (e.g. dol m 1 and dol m 1.01, lep d 2.0102 and lep d 2, per a 1 and per a 1.0103, equ c 2 and equ c 2.0102, etc.) which mainly contain numbers in their names. In the case of “Proteins Name”, the algorithm manages to find only a small percentage of the target variants (19 out of 79). It recognized incorrectly as variants certain strings that were containing redundant specifications and should not be included in the ontology. For example, the word “manganese” is redundant in “manganese superoxide dismutase” and “manganese-superoxide dismutase” but the algorithm groups them with the correct instance “superoxide dismutase”. The same happens with the word “vacuolara” that appears in “vacuolar serine protease” and “serine

protease”. This happens because the algorithm is misled by the extra characters the string contains. The above is a strong evidence towards the need of optimizing the performance of the information extraction engine used before, since the incorrect identification of instances affects the performance of COCLU.

7 Concluding Remarks

In this work we presented our methodology for building a formally defined ontology, incrementally maintaining it exploiting machine learning techniques and domain specific corpora, and evaluating it using a well defined experimental setting. We examined the application of this methodology in the biomedical domain of allergens:

- designing and implementing a formally defined ontology on allergens exploiting existing taxonomies (semi-formal models) and documents that describe the allergen nomenclature,
- maintaining semi-automatically the allergens ontology discovering knowledge (instances and typographic variants of instances) from domain specific corpora (PubMed abstracts on allergens) using machine learning techniques,
- evaluating the maintenance process through the study of the factors that may affect its performance (size of the initial allergens ontology, structure of the HMM-based extraction technique in the knowledge discovery stage, parameters of the clustering algorithm COCLU in the knowledge refinement stage).

The initial results are quite encouraging. The coverage of the ontology increased to 81.9% in two iterations starting from a coverage of only 34.9%. Also, the clustering algorithm COCLU which was used to discover typographic variants of existing ontological instances performed quite well in the case of “Allergen Sources Name” variants. Its rather poor performance in the case of “Allergens Scientific Name” and “Proteins Name” variants, was mainly due to domain specific peculiarities and incorrectly extracted instances. We believe that its performance can be further improved tuning the COCLU threshold parameter and enabling the use of domain specific features. Concluding, the combination of the ontology-based and the machine learning based annotation methods gave good results on a corpus of unstructured content exploiting only token type information.

To sum up the presentation of the current status of our work, we would like to point out the major innovative aspects:

- The application in the allergens domain enabled us to refine our ontology maintenance process and integrate it into a methodology for the whole process of ontology development. This methodology is appropriate for evolving domains where the domain knowledge must be regularly updated.
- We developed a formally defined ontology on allergens following well established design principles properly adapted to the domain requirements. This ontology was implemented in a description logic based language (OWL) in order to facilitate its re-use by other researchers in the areas of ontologies and allergens. This ontology will become publicly available soon.
- We specified design paradigms for the building of formally defined ontologies following the well accepted stages of ontology design and enforcing generally accepted design criteria properly adapted to the allergen domain. This experience can be useful to other relevant areas in biomedicine.
- The initial ontology used in our experiments was derived from a part of the allergen corpus containing abstracts published up to a certain point in time. This was done in order to simulate the evolution process for a new domain. The use of a richer formally defined ontology, which is not derived from a specific corpus but from existing allergens or databases, as a starting point may lead to practical applications with even better results.
- We introduced a new lexico-semantic relation, quite important in evolving domains where new names and variants are continuously added, and examined the ability of our clustering algorithm in locating this relation.
- We specified an evaluation methodology which describes the stages followed and examines the factors that may affect the system performance.

Based on the experiences collected so far, we will continue our work in order to improve further the proposed approach. More specifically, our plans include the following:

- Application of the complete methodology in at least one more domain.
- Document the allergens ontology and make it publicly available.
- Examine the application of at least one more information extraction technique in the knowledge discovery stage and compare its performance with the HMM-based one that is currently being used.
- Experiment further with COCLU in the *'hasTypographicVariant'* relation adding domain specific features and tuning the threshold parameter. We will also examine its application for the identification of other relations.
- Study the evolution rate of a domain in order to provide hints on the proper period in time in which the proposed methodology should be applied again.
- Develop techniques for grouping properties that belong to the same entity. Some initial work on this will be presented in the AIME'05 conference [44].
- Apply our methodology to full-text articles instead of abstracts towards large-scale experiments.
- Develop a platform that will integrate all the tools we are using in order to facilitate the application of our methodology.

Concluding, we believe that the route to advanced intelligent techniques passes through the formal modeling of various domains of interests using ontologies, therefore their development and maintenance should be methodologically supported to secure high standards.

Acknowledgements

The authors are grateful to Prof. Brusica and his team at the Institute for Infocomm. Research, Singapore, for providing the PubMed abstracts on allergens and commenting on various stages of this work. Thanks also go to Dr. S.Konstantopoulos and Mr. E.Fritzilas for helping us with their comments on the paper structure and the use of English. Finally, we would like to thank the anonymous reviewers for their valuable suggestions and comments.

References

- [1] D.L. McGuinness, *Ontologies Come of Age*, in: Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential, edited by D. Fensel, J. Hendler, H. Lieberman, and W. Wahlster, MIT Press, pp. 171-194, 2003.
- [2] A. Valarakos, G. Paliouras, V. Karkaletsis, and G. Vouros, *Enhancing Ontological Knowledge through Ontology Population and Enrichment*, in: Proceedings of the 14th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2004), LNAI, Springer-Verlag, vol. 3257, pp. 144-156, 2004.
- [3] N.F. Noy and M. Klein, *Ontology Evolution: Not the Same as Schema Evolution*, in: Knowledge and Information Systems, 6:428-440, 2004.
- [4] H.S. Pinto and J.P. Martins, *Ontologies: How can they built?*, Knowledge and Information Systems (KAIS), Springer-Verlag, 6:441-464, 2004.
- [5] D. Jones, T. Bench-Capon, and P. Visser, *Methodologies for ontology development*, in: Proceedings of the IT&KNOWS Conference, XV IFIP World Computer Congress, Budapest, 1998.
- [6] M. Fernandez Lopez, *Overview of Methodologies for Building Ontologies*, in: Proceedings of the Workshop on Ontologies and Problem-Solving Methods: Lessons Learned and Future Trends (IJCAI99), Stockholm, 1999.
- [7] M. Craven, D. DiPasquo, D. Freitag et al., *Learning to Construct Knowledge Bases from the World Wide Web*, Journal of Artificial Intelligence, vol. 118, no. 1/2, pp. 69-113, 2000.
- [8] O. Etzioni, S. Kok, Soderland et al., *Web-Scale Information Extraction in KnowItAll (Preliminary Results)*, in: Proceedings of the 13th International World Wide Web conference (www2004), pp. 100-110, 2004.
- [9] M.A. Hearst, *Automatic Acquisition of Hyponyms from large Text Corpora*, in: Proceedings of the 14th International Conference on Computational Linguistic (COLING), vol. 2, pp. 539-545, 1992.
- [10] M. Vargas-Vera, J. Domingue, M. Lanzoni, A. Stutt, and F. Ciravegna, *MnM: Ontology driven semi-automatic support for semantic markup*, in: Knowledge Engineering and Knowledge Management (Ontologies

- and the Semantic Web), edited by A. Gomez-Perez, and V. R. Benjamins, Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management (EKAW02), LNAI, Springer-Verlag, vol. 2473, 2002.
- [11] A. Harith, K. Sanghee, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, and N. Shadbolt, *Web based knowledge extraction and consolidation for automatic ontology instantiation*, in: Proceedings of the Workshop on Knowledge Markup and Semantic Annotation (KCap'03), Florida, USA, 2003.
- [12] F. Ciravegna, A. Dingli, D. Guthrie, and Y. Wilks, *Integrating Information to Bootstrap Information Extraction from Web Sites*, in: Proceedings of the IJCAI Workshop on Information Integration on the Web, pp. 9-14, 2003.
- [13] A. Kiryakov, B. Popov, D. Ognyanoff, D. Manov, A. Kirilov, and M. Goranov, *Semantic Annotation, Indexing, and Retrieval*, in: Proceedings of the 2nd International Semantic Web Conference (ISWC2003), Florida, USA, LNAI, Springer-Verlag, vol. 2870, pp. 484-499, 2003.
- [14] M.A. Hearst, *Automated Discovery of WordNet Relations*, in: C. Fellbaum (Ed.), *WordNet: An Electronic Lexical Database*, MIT Press, pp. 132-152, 1998.
- [15] C. Brewster, F. Ciravegna, and Y. Wilks, *User-Centered Ontology Learning for Knowledge Management*, in: Proceedings of the 7th International Conference on Applications of Natural Language to Information Systems, LNCS, Springer-Verlag, vol. 2553, pp. 203-207, 2002.
- [16] P. Buitelaar, D. Olejnik, and M. Sintek, *A protégé plug-in for ontology extraction from text based on linguistic analysis*, in: Proceedings of the First European Semantic Web Symposium, (ESWS), edited by C. Busler, J. Davies, D. Fensel, and R. Studer, LNCS, Springer-Verlag, vol. 3053, pp. 31-44, 2004.
- [17] U. Hahn and K.G. Markó, *An integrated, dual learner for grammars and ontologies*, in: *Data & Knowledge Engineering*, 42(3):273-291, 2002.
- [18] G. de Chalendar and B. Grau, *SVETLAN or how to Classify Words using their Context*, in: Proceedings of the 12th European Workshop on Knowledge Acquisition, Modeling and Management, LNCS, Springer-Verlag, vol. 1937, pp. 203-216, 2000.

- [19] D. Faure and C. Nedellec, *Knowledge acquisition of predicate argument structures from technical texts using machine learning: The system asium*, in: Proceedings of the 11th European Workshop on Knowledge Acquisition, Modeling and Management, LNCS, Springer-Verlag, vol. 1621, pp. 329-334, 1999.
- [20] A. Maedche and S. Staab, *Discovering conceptual relations from text*, in: Proceedings of the ECAI-2000, IOS Press, Amsterdam, pp. 321-325, 2000.
- [21] T. Yamaguchi, *Acquiring Conceptual Relationships from Domain-Specific Texts*, in: Proceedings of the Second Workshop on Ontology Learning OL'2001, Seattle, 2001.
- [22] A. Faatz and R. Steinmetz, *Ontology Enrichment with Texts from the WWW*, in: Proceedings of the 2nd International Workshop on the Semantic Web, Finland, 2002.
- [23] H. Wache, T. Vögele, U. Visser, H. Stuckenschmidt, G. Schuster, H. Neumann and S. Hübner, *Ontology-Based Integration of Information - A Survey of Existing Approaches*, in: Proceedings of the IJCAI-01 Workshop: Ontologies and Information Sharing, Seattle, WA, pp. 108-117, 2001.
- [24] P.G. Baker, C.A. Goble, S. Bechhofer, N.W. Paton, R. Stevens, and A Brass, *An Ontology for Bioinformatics Applications*, Bioinformatics, 15(6), pp. 510-520, 1999.
- [25] Gene Ontology WWW resources: <http://www.geneontology.org> (new project in owl)
- [26] S. Schulze-Kremer, *Ontologies for Molecular Biology*, in: Proceedings of the 3rd Pacific Symposium on Biocomputing, World Scientific Publishers, pp. 693-704, 1998.
- [27] CSO ontology WWW resources:<http://ontology.ims.u-tokyo.ac.jp/signalontology/>
- [28] V. Giudicelli and M.P. Lefranc, *Ontology for immunogenetics: the IMGT-ONTOLOGY*, BIOINFORMATICS, Oxford University Press, vol. 15, no. 12, pp. 1047-1054, 1999.
- [29] V. Brusic, M. Millot, N. Petrovsky, S.M. Gendel, O. Gigonzac, and S.J. Stelman, *Allergen databases*, Allergy, 58(11):1093-1100, 2003.

- [30] T. Gruber, *It Is What It Does: The Pragmatics of Ontology*, Invited presentation to the meeting of the CIDOC Conceptual Reference Model committee, Smithsonian Museum, Washington, D.C., March 26, 2003.
- [31] T. Gruber, *Towards Principles for the Design of Ontologies Used for Knowledge Sharing*, International Journal of Human-Computer Studies, 43:907-928, 1995.
- [32] T. Bittner, *Axioms for parthood and containment relations in bio-ontologies*, in: Proceedings of the Workshop on Formal Biomedical Knowledge Representation (KR-MED), pp. 4-11, 2004.
- [33] <http://www.w3.org/TR/owl-ref/>
- [34] G. Zhou, J. Zhang, Z. Su, D. Shen, and C. Tan, *Recognizing Names in Biomedical Texts: a Machine Learning Approach*, Bioinformatics, Oxford University Press, 20(7):1178-1190, 2004.
- [35] D. Hanisch, J. Fluck, H.T. Mevissen, and R. Zimmer, *Playing Biology's name game: Identifying protein names in scientific text*, in: Proceedings of the Pac Symp. Biocomput., pp. 403-414, 2003.
- [36] A. Valarakos, G. Paliouras, V. Karkaletsis, and G. Vouros, *A Name-Matching Algorithm for Supporting Ontology Enrichment*, in: Proceedings of the Hellenic Conference in Artificial Intelligence (SETN), LNAI, Springer-Verlag, vol. 3025, pp. 381-389, 2004.
- [37] V. Karkaletsis, C.D. Spyropoulos, C. Grover, M.T. Paziienza, J. Coch, and D. Souffis, *A Platform for Crosslingual, Domain and User Adaptive Web Information Extraction*, in: Proceedings of the European Conference in AI, pp. 725-729, 2004.
- [38] Message Understanding Conference: http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html
- [39] S. Boisen, M. Crystal, R. Schwartz, R. Stone, and R. Weischedel, *Annotating Resources for Information Extraction*, in: Proceedings of the 2nd International Conference on Language Resources & Evaluation, pp. 263-266, 2000.
- [40] Protege owl plug-in: <http://protege.stanford.edu/plugins/owl/>
- [41] D. Freitag and A. McCallum, *Information extraction using hmms and shrinkage*, in: Proceedings of the Workshop on Machine Learning for Information Extraction (AAAI-99), pp. 31-36, 1999.

- [42] K. Seymore, A. McCallum, and R. Rosenfeld, *Learning hidden markov model structure for information extraction*, Journal of Intelligent Information Systems 8:5-28, 1999.
- [43] G. Petasis, V. Karkaletsis, G. Paliouras, I. Androutsopoulos, and C.D. Spyropoulos, *Ellogon: A New Text Engineering Platform*, in: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), pp. 72-78, 2002.
- [44] A. Valarakos, V. Karkaletsis, D. Alexopoulou, E. Papadimitriou, and C.D. Spyropoulos, *Populating an Allergens Ontology using Natural Language Processing and Machine Learning Techniques*, to appear In Proceedings of the 10th Conference on Artificial Intelligence in Medicine (AIME 05), 2005.