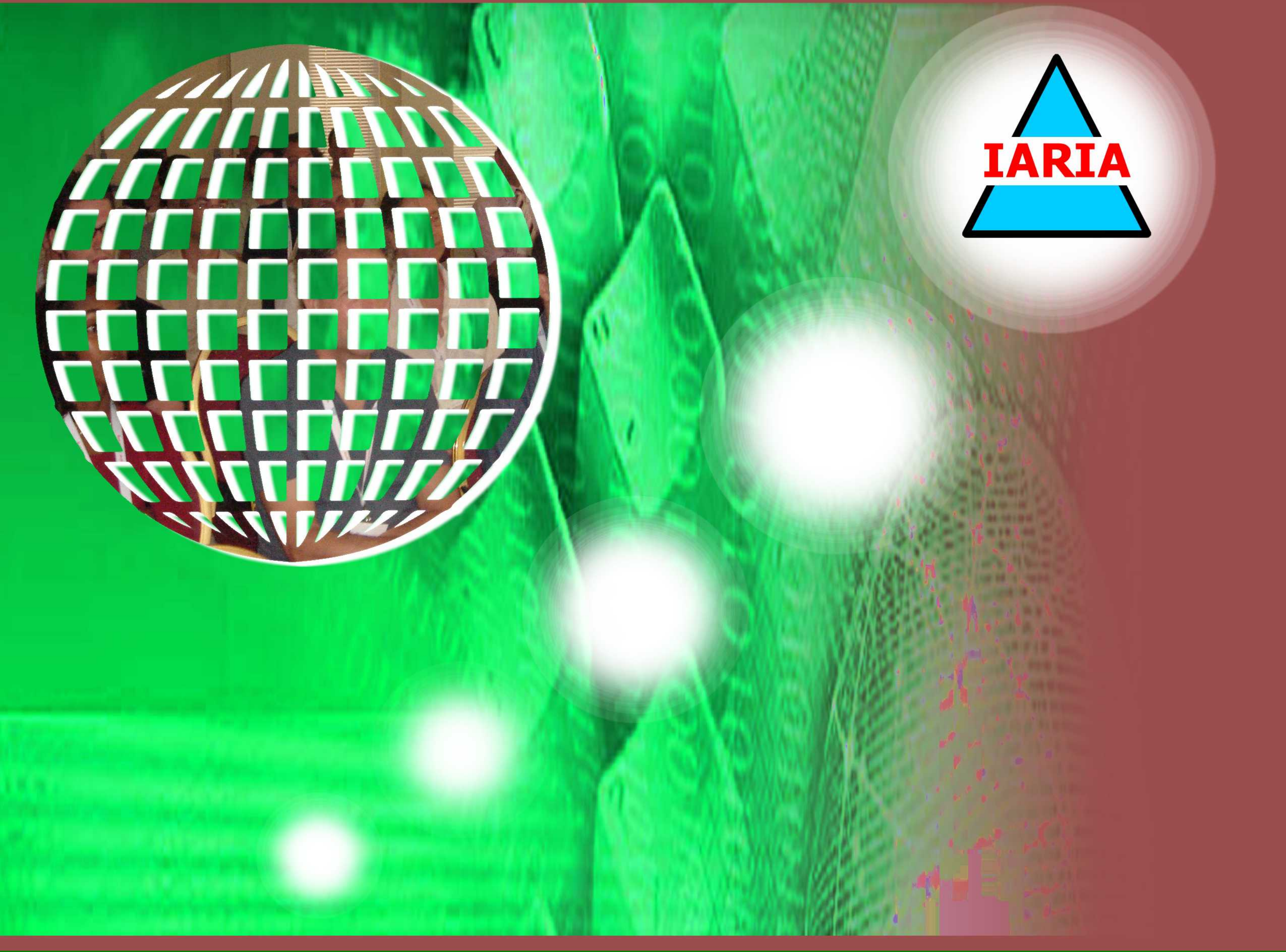
International Journal on Advances in Life Sciences



2009 vol. 1 nr. 2&3

The *International Journal on Advances in Life Sciences* is Published by IARIA.
ISSN: 1942-2660
journals site: http://www.iariajournals.org
contact: petre@iaria.org

Sponsored by IARIA
www.iaria.org

International Journal on Advances in Life Sciences
Volume 1, Numbers 2&3, 2009

Editor-in-Chief

Editorial Advisory Board

EHealth and eTelemedicine

Electronic and Mobile Learning

Advanced Knowledge Representation and Processing

International Journal on Advances in Life Sciences
Volume 1, Numbers 2&3, 2009

CONTENTS

Towards an Ontology of Crucial Knowledge Identification to Improve the K-DSS

Sabine Bruaux
Laboratory of Modelisation, Information, System
Picardie Jules Verne University
Amiens School of Management
Amiens, France
sabine.bruaux@u-picardie.fr

Inès Saad
Laboratory of Modelisation, Information, System
Picardie Jules Verne University
Amiens School of Management
Amiens, France
ines.saad@u-picardie.fr

***Abstract*— In this paper, we propose a characterization of the main classes contained in the database of the system K-DSS and related to the domain of identification of the crucial knowledge for which a capitalizing operation is required. We exploit ontological categories existing in the literature to define the notions of Knowledge, Actor, Support and Criteria of knowledge vulnerability. The objective is to improve the process of the crucial knowledge evaluation providing to different decision makers a unified semantic of these entities. Such approach brings us a preliminary analysis for the construction of an application ontology that we aim to integrate as a new component of the decision support system K-DSS.**

Keyword: Knowledge management; Ontology, Crucial Knowlegde, K-DSS, Multi-criteria Decision Aid

I. INTRODUCTION

The necessity to create and to use knowledge mobilized and produced in firms has increased rapidly these last years. Firms become aware of the importance of the immaterial capital owned by their employees which corresponds to their experience and accumulated knowledge about the firm activities. Maintaining this capital is powerful mean to improve the level of performance of the firm. In order to create, preserve and share knowledge in firms, Knowledge Management has been occupying since the beginning of the nineties a more and more important place within organizations. Thus, companies should invest in engineering methods and tools [11] in order to preserve knowledge especially those of *tacit* nature. Researchers in knowledge engineering and knowledge management have been focusing on the problems of acquisition, preservation and transfer of knowledge. However, considering the large amount of knowledge to be preserved, the firm must first determine knowledge that should make the object of capitalization. We should focalize on only the so called “crucial knowledge”, i.e. the risk of their lost and the cost of their (re)creation is considered important; their contribution to reach the project objectives is very important and their use duration is long. Our previous research works also revealed the interest of the identification of crucial knowledge [34]. Not enough works exist concerning the identification of knowledge on which it is necessary to capitalize [18] [34] [41]. Thus, we have proposed a multicriteria method based on dominance rough set approach to identify and qualify crucial knowledge in order to justify a situation where knowledge capitalization is advisable. The value added of our methodology is to elicit the preference of the decision makers. The proposed method was conceived and validated in the French car Company [33]. This method is supported by a decision support system called K-DSS [34]. Our system K-DSS is based on two types of tasks: automation task and human task.

The K-DSS system implements a database in the form of a UML diagram of classes, which models the process of the knowledge assessment on a criteria family. However, this database has been designed without to give some meaning to the classes that it contains (e.g., the classes of *knowledge*, *process*, *actor*) [8]. Currently, the different criteria of evaluation (e.g., *scarcity*, *complexity*, *portability*) are the attributes of class knowledge. However, the notion of knowledge doesn’t need the notions of scarcity, complexity or portability to be defined. We think that lack of semantic

In order to improve the performance of K-DSS, a first work is to specify the semantics of the UML classes in an ontology of the domain of the potentially crucial knowledge assessment. The construction of a such ontology in the context of the knowledge management system K-DSS, will define a shared vocabulary about the knowledge evaluation on the vulnerability criteria. This involves to define the elements of knowledge to which it is referred in the database (such as knowledge, tacit knowledge, explicit knowledge, individual knowledge, collective knowledge, actors, criteria of vulnerability, etc.), the relations between these elements and the semantics that they should be interpreted.

This article presents the first step of our process of the construction of an ontology reflecting the process of the knowledge potentially crucial evaluation.

In the following sections of this article, we first present the functional architecture of the decision support system K-DSS. We describe in particular the UML classes and relations involved in the process of identification of the knowledge (section 2). In the section 3, we expose a literature review to explain and justify our methodology of the construction of an ontology covering the domain of the evaluation of crucial knowledge. Then, we present the result of the first phase of the construction of an ontology, which means the conceptualization of the domain of the evaluation of crucial knowledge. We define in an informal language the concepts of Knowledge, Actor, Support and Criteria (section 4). Finally, we present our conclusions and perspectives (section 5).

II. Related Work

The Knowledge-Based Systems (KBS) are defined to reuse and share all or parts of the knowledge bases in order to extend the class of problems to be solved (e.g., car repairs, medical diagnostics, etc..) or to rely on skills of other systems (e.g., obtain an advice about a rare disease).

"Building knowledge-based systems today usually entails constructing new knowledge bases from scratch. It could be done by assembling reusable components. System developers would then only need to worry about creating the specialized knowledge and reasoners new to the specific task of their system. This new system would interoperate with existing systems, using them to perform some of its reasoning. In this way, declarative knowledge, problem-solving techniques and reasoning services would all be shared among systems. This approach would facilitate building bigger and better systems cheaply." [39]

The principle of KBS relies on its internal structure. Since the middle of 80s, the modeling of knowledge for the development of a KBS differentiates the representation of the terminological knowledge of a domain from the modeling of treatments we want done on these knowledge, that we call the inferential knowledge. Basically, different kinds of knowledge are exploited by the KBS:

- the domain knowledge. For example, the knowledge: "a meningitis is common severe headaches" focuses on the disease meningitis, it helps to define by precising one of its frequent manifestation[1].
- the control knowledge, which details a method of use of the domain knowledge to solve a problem. For example, the knowledge: "if the patient has a sign corresponding to the frequent manifestation of a disease, then mention this disease as a hypothesis of diagnosis" exploits the facts to provide a method and discuss the hypothesis of diagnosis.
- the rules, which are formulated in the form of empirical associations between the characteristics of the problem and the possible solutions. For example, the knowledge: "if the presence of severe headaches, then think about meningitis" domain knowledge and control knowledge.
- the constraints, which allow to specify impossibilities or obligations, for example: "a meningitis can affect anyone at any age, from newborns to seniors".

Thus, the KBS is able to solve a problem through a series of deductions (inferences). The construction of the KBS has asked the question of the construction of models of knowledge (e.g., domain model, reasoning model) and has highlighted the fact that during this construction takes place a process of creation of new knowledge. Such knowledge is not a knowledge already present in the head of the expert (s). The term "knowledge-based system" reflects this new approach of knowledge acquisition. Thus, from the initial practice of knowledge acquisition (do precisely the reasoning human), we have moved progressively to a practice of the structuration and the formalization of knowledge, in other words, a practice of the construction of models.

The researches are oriented now to the activity of the construction of a model of knowledge, which is no more focused on the problem-solving performance of the system (similar to those of the expert), but on how the problem-solving knowledge are used in interaction with the user into the cooperative systems. The research on the sharing, the reuse of knowledge bases and the semantic interoperability of KBS (ARPA Knowledge Sharing Effort project [39]) requires the use of ontologies to express knowledge by using the primitive of specification defined at a conceptual level, independently of any formal representation.

An ontology is an explicit specification of a conceptualization covering a domain of knowledge [17]. The term "explicit specification" means that the design is represented in a natural language or formal. The term "conceptualization" refers to a system of concepts. An ontology defines the central terms of a domain of knowledge and the consensual semantics associated with these terms, in the form of concepts related to each other by taxonomic (hierarchical) and semantic relations [42]. In the medical domain, for example, the knowledge will focus on the function of an organ, the effects of an antibiotic, or the manifestations of a disease. The domain is divided into categories of entities such as: "body", "pathophysiological process," "disease," "physiological function," linked by relations: "causes", "manifested by", "provides the function of ".

A concept can be defined as an entity composed of a term (e.g., the term "star"), an intension, which is the set of properties reflecting the meaning of the concept (e.g., a bright spot in the sky at night) and an extension that is the set of the objects (called instances of the concept) denoted by the concept (all bright spots). This method of definition is a long tradition that can be traced back to the Greek philosopher Aristotle [4]. By convention, a relation is also characterized by a term (e.g., "to be the author of"), an intension, which helps to express the meaning of the relation by specifying the concepts that it connects (e.g., "R is a relation between a person or a group who created a document, and its intellectual content, its arrangement or its shape") and an extension that is the set of the tuples of instances linked by the relation (e.g.: (Hugo, Notre Dame Paris)). The relations have in addition a "signature", a list specifying the types of instances that they connect, or for our example: (person, Document).

We have seen that the properties (or intensions) of concepts and relations involved in the definition of the semantics of a domain of knowledge. More generally, all the properties specific to the domain of knowledge, which

[1] We take the example of a KBS used to generate diagnostic hypotheses proposed in [23].

contribute to express the meaning of concepts and relations, and how to use them in the application, are represented in the ontology by axioms. Since the objective is to share meaning, these primitives should get the meaning of the term as objectively as possible, i.e., independently of the use that we want to do of these knowledge [21]. "In order to integrate an ontology in a KBS, it should be translating into a form suitable for the use of the KBS, i.e., it should be specified the semantic of the manipulation of axioms. Thus, an axiom can be used to infer new knowledge or to validate the adequacy of a knowledge in relation to the semantic of the domain" [29].

As the ontologies are particular knowledge bases, the methods of the construction of the ontologies incorporate the main principles of the construction of the KBS. In particular, the construction of an ontology is done by successive transformations of ontological models. It is customary to distinguish three main stages in the construction of an ontology in a formalism that allows the manipulation of knowledge in an domain with a KBS [29]:

1. **The knowledge acquisition.** This process consists in identifying a corpus (which may contain for example terminological bases, technical documentation, summaries of interviews, questionnaires) covering all the documents of a given domain, knowledge for the operational needs in terms of concepts, relations, instances and axioms (i.e., the semantics of the domain). This process, which, from raw data, leads to a conceptual model informal (e.g., in a natural language) or semi- informal (ex : CML [35] and UFO [17] languages), is called conceptualization.
2. **The knowledge modeling.** This is to structure all conceptual entities, identified during the step of acquiring knowledge, and to formalize them in a language of representation of ontologies (e.g., the languages based on frames, the description logics, Conceptual Graphs [36], Ontolingua [19], RDF [22]). This process, which, from a conceptual model leads to an ontology (semi-formal) is called ontologization.
3. **The knowledge representation.** This is to clarify the semantic of the manipulation of the axioms in order to allow the KBS to reason about the knowledge of the domain (depending on the scenario for use by the application, like enable the KBS to take decisions). The ontology obtained in the previous step must therefore be specified in an operational representation (e.g., FLogic, KIF, OCML, RDFS, DAML, OIL, OWL), i.e., a formal language that has inferential mechanisms (facts, rules and constraints). This process of specification of a semi-formal ontology into a model executable by a machine (operational ontology or computational ontology) is called operationalization.

Several methodologies have been proposed for building ontologies. Some methodologies are planning to take over the whole process of the specification at the conceptual level of an ontology to its formalization (e.g., METHONTOLOGY [21], TERMINAE [17], the method of Gruninger and Fox [19], On-To-Knowledge [37]). Thus, they distinguish two levels of modeling: the modeling to establish the meaning and the modeling to implement a KBS. Other methods focus on one phase of the process (conceptualization, ontologization, operationalization). The methods Cyc [44], SENSUS [40], the approach KACTUS [27] and the method of Uschold and King [42] for example insist on the stage of conceptualization. The methods OntoSpec [27], Archonte [29] and OntoClean [22] provide a help for the structuration of the hierarchies of concepts and relations during the phase of ontologization.

Like the methodologies, many tools to build the ontologies have been developed. These include: KAON [39], OntoEdit [39] based on the methodology On-To-Knowledge, Protégé-2000 [30], Oiled [5], WebODE [1] that implements the methodology METHONTOLOGY.

The construction of a KBS begins, before the implementation (not to constraint the representations with the criteria of performance or computability), by the abstract description of the system, by using the primitive of specification of the conceptual model at the knowledge level (KADS method [23], method MACAO method [33]). It is possible to take advantage of the existence of the repeated structures in the conceptual models. The reuse of the generic components is a process of specialization which consists in adapting the generic problem-solving method the most appropriate to a class of problems in the application domain. The model of reasoning of the application is a specialization of the generic problem-solving methods selected. This principle can also be applied to the elements of a domain, by reusing a generic domain ontology that contains generic concepts from the systemic (e.g., state, function, system). A help to the modeling of knowledge consists in reusing ontologies already built.

In the next section, we present the result of the first phase of the construction of an ontology relative to the process of crucial knowledge evaluation. The phase of the conceptualization of the domain of crucial knowledge evaluation is the most long and the most difficult. This phase consists in identifying the terms structuring the domain of the potentially crucial knowledge evaluation in terms of concepts, relations, instances and axioms (e;g., define the minimum and sufficient conditions to to say that an object belongs to a given class), from available resources. Here, the resources are the interviews with experts modeled as a diagram of UML class [33]. This diagram models the process of critical knowledge evaluation. We apply a "middle-out" approach for the identification of the central concepts of the ontology, we will generalize and specialize to complete the ontology [43]. It is recognized that this approach promotes the modularity and the stability of the resulting ontology. We also exploit the ontological frameworks existing in the literature (parts of high-level ontologies and domain ontologies) to clarify the definitions of concepts and relations of ontology.

III. Functional architecture of K-DSS

This section describes the functional architecture of the system K-DSS.

First, it is important to identify the specialized roles played by the persons concerned by the knowledge identification decision system (Figure 1).

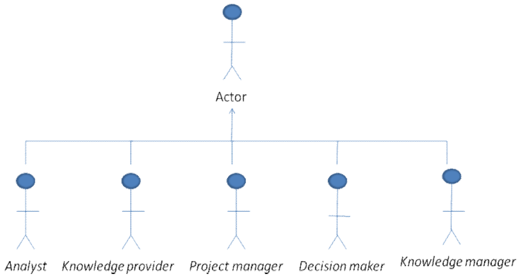


Figure 1. the specialized roles played by the persons concerned by the knowledge identification decision system

The following list enumerates the main involved internal and/or external actors of the organization:

- *Knowledge provider*. An important role in the crucial knowledge identification decision process is played by the knowledge "owner" or provider. The knowledge provider is generally an expert in the project under study but can also be a different person in the organization who is not considered to be an "expert" .
- *Project manager*. The project manager is responsible for running the project considered by the crucial knowledge identification decision process. So, she or he is involved in all the phases of the decision process.
- *Decision maker*. A decision maker is an individual or a group of individuals who, because of their value system, directly influence the final recommendation.
- *Knowledge manager*. The knowledge manager formulates knowledge identification, preservation, distribution and actualisation.
- *Analyst*. An analyst is not involved in development project. He formulates criteria and preference model to help decision makers for using the system and identifying crucial knowledge.

Two phases may be distinguished. The first phase is relative to the construction of the preference model. The preference model is represented in terms of decision rules. The second phase concerns the classification of "potential crucial knowledge" by using the rules collectively identified by all the decision makers during the construction of the preference model.

A. *Construction of the preference model*

This phase consists in identifying, from the ones proposed, an algorithm for computing the contribution degrees. The selection is collectively established by all the decision makers with the help of the analyst. Whatever the selected algorithm, it uses the matrices Knowledge-Process (K-P), Process-pRoject (P-R) and pRoject-Objective (R- O) extracted from the database more specifically from the three association classes "Evaluate-K-P", "Evaluate-P-R "and "Evaluate-R-O " to compute the contribution degree of each piece of knowledge into each objective. To avoid data redundancy, these matrices are not explicitly stored in the database but generated during processing. Only their intentional definitions are permanently stored in the system.

Once these matrices are generated, the contribution degrees are first stored temporally in a decision table and then introduced in the database. As for matrices, only the intentional definition of the decision table is maintained in the system.

The decision table (Table 1) contains also the evaluation of the "Reference crucial knowledge" concerning the vulnerability and use duration criteria extracted from the class "Knowledge" precisely. These evaluations are collectively defined and introduced by the analyst into the database. The analyst should introduce in the decision table, and for each decision maker, the decisions concerning the assignment of "Reference crucial knowledge" into decision classes Cl1: "Not crucial knowledge" and Cl2:"Crucial Knowledge".

K_i	g_1	g_2	g_3	g_4	g_5	g_6	g_7	g_8	g_9	g_{10}	g_{11}	g_{12}	g_{13}	g_{14}	g_{15}	Decision class
K_8	2	2	3	3	1	2	4	4	5	2	4	5	5	5	2	Cl1
K_9	3	3	2	2	3	3	4	4	4	2	4	4	3	4	2	Cl1
K_{16}	2	3	3	2	2	2	3	4	5	2	5	5	5	2	2	Cl2

Table 1. An extraction from the decision table for one decision maker

The decision table contains, in addition to the columns relative to vulnerability and those relative to contribution degree and use duration criteria, as many columns as decision makers. Once the decision table is generated, it will be used as the input of the induction algorithm selected by the decision makers.

This algorithm permits to generate the list of the initial rules for each decision maker. It is important to mention again that only rules relative to class Cl2 are stored. Then, each decision maker should select a subset from these initial rules. The next step in this phase consists to collectively select, from the set of decision rules individually identified by the different decision makers, a subset of decision rules that will be used latter by JESS for the classification phase.

B. *Evaluation of "potential crucial knowledge"*

The second phase consists in classifying the new knowledge called "potential crucial knowledge". As the previous one, this phase starts by identifying the algorithm to be used for computing the contribution degree of each piece of knowledge into each objective. This algorithm uses as input the information relative to the performances of "potential crucial knowledge" previously introduced in the matrices K-P, P-R and R-O. The results are stored in a performance table. The information contained in the performance table are then transformed into facts. The inference engine incorporated in JESS verifies first if exists at least one rule that verifies the different facts and if this holds, the piece of knowledge is classified as crucial; otherwise the piece of knowledge is considered non crucial.

C. Database

The UML-based conceptual schema of the database is shown in figure 3. The central class in the model is the class "Knowledge". It is described with an unique number (K-Num), a name (K-Name), a description (K-Description), eight attributes (Complexity-Level, Substitutability-Level, Validation-Level, Transferability-level, Scarcity-Level, Acquisition-Cost, Production-Time, and Accessibility-Level) corresponding to the eight criteria g_1,g_2,…, and g_8 composing knowledge vulnerability family, use duration (Use-Duration) corresponding to the only criterion, g_{15}, of use duration family, (Knowledge-Type) (i.e. "reference crucial knowledge" or "potential crucial knowledge").

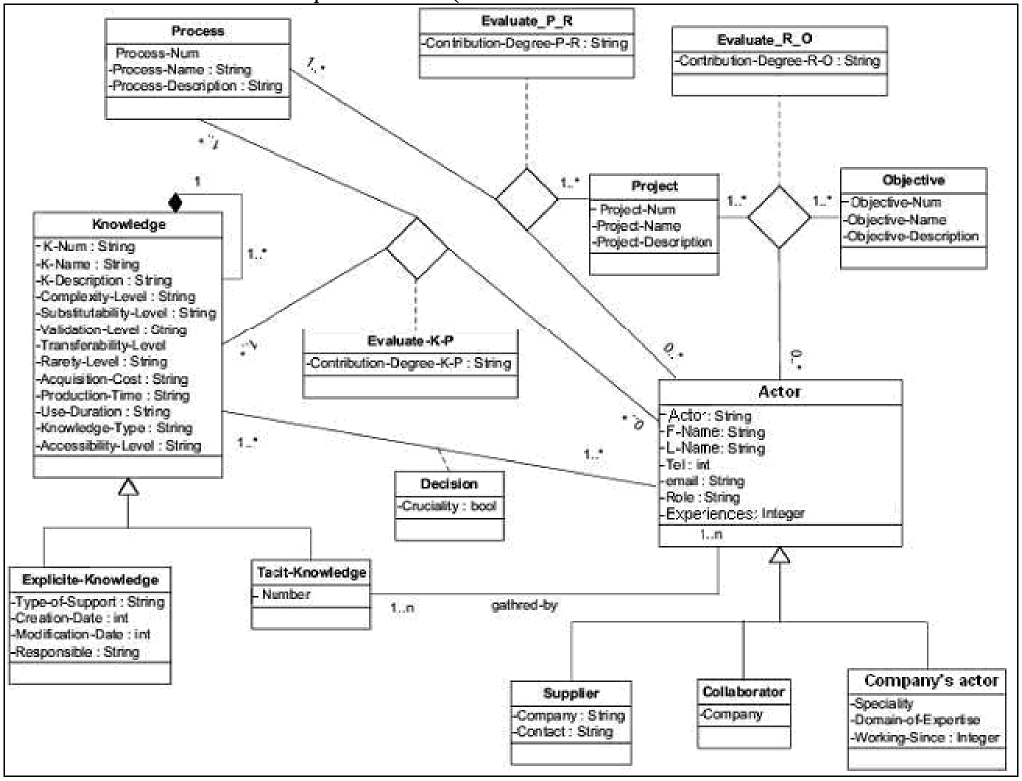


Figure 3. Database [Saad, 2005]

Below we quickly specify the content of the criteria used. These criteria are constructed based on a real context and a real-world case study conducted in an automobile company. We believe that these criteria are generally valid for the entire problem of identification of knowledge requires an operation funded through a transfer to similar projects[2]:

- *Scarcity* represents the number of person (internal or external to the organization) who own the knowledge.
- *Transferability* is the degree of the transfer of individual or collective knowledge. We have based our analyze on the definition given by Davenport and Prusak [10] "*knowledge transfer involves two actions: transmission and absorption by that person or group*" to measure the degree of transferability of knowledge. So we distinguish two states for measuring the knowledge transferability :
    - o *Transmission* represents the degree to which an individual or group of individual can transmit his knowledge to other person. It is difficult to transmit individual knowledge because it is fundamentally tacit. Knowledge incorporates so much embedded learning that its rules may be separate from how individual acts.
    - o *Absorption* is the degree to which individual can appropriate the knowledge by either studying technical document or talking to her predecessor skilled individual.

[2] So far, we suggest to the knowledge manager an update of the definitions of criteria and scales if necessary, according to the needs of actors.

- *Imitability* represents the degree to which competitors can copy knowledge by analyzing patent, by experiencing product, etc.
- *Accessibility* represents the time needed to access to the knowledge. The accessibility notion is relative because it has to be compared with the length of time the person actually has to access to the knowledge.
- *Complexity* represents the degree to which multiple kind of knowledge domain are needed to create a knowledge in the process or to adapt it to another context.
- *Validity* represents the validation state of knowledge. We distinguish two types of validation :
  - knowledge validated by macroscopic or microscope experiments
  - knowledge validated by experts ; for example; thesis, patent...
- *Substitutability* is the degree to which knowledge can be replaced by an other knowledge to take the same task with the same performance.
- *Cost and time of knowledge production* represents the number of persons and the period needed to create the knowledge.
- *Use duration :* The evaluation concerning criterion g_{15} is provided by experts. For example, the knowledge relative to "the measurement of the additive" has an "average use duration" because is related to the use duration of the first generation of depollution system; new generations of the depollution systems are without additive.

Note finally that a piece of knowledge may be composed of several elementary pieces of knowledge. This is modelled by the aggregation relation defined on the class "Knowledge".

The classes "Explicit-Knowledge" and "Tacit-Knowledge" are specializations of the class "Knowledge". The "Explicit- Knowledge" class permits to identify for each explicit knowledge the set of supports (documents, database, knowledge base system) on which this knowledge is represented. If the knowledge is tacit, it is characterized with the person who gathers it. This information is deduced from the relationship between "Tacit-Knowledge" and "Actor". The class "Actor" contains the information relative to the different actors (Id, Name, Telephone, Email, Role, Experience). The class "Actor" is specialized into three classes: "Supplier", "Collaborator" and "Company's Actor".

The three classes: "Process", "Project" and "Objective" permit to handle the information relative to the names and descriptions of processes, projects and objectives, respectively. The association class "Evaluate-K-P" between "Actor", "Process" and "Knowledge" stores the contribution degree of a knowledge into a process (Contribution-Degree-K-P) attributed by a given actor.

IV. Conceptualization of the domain of crucial knowledge evaluation

The various experiments "in situ" revealed that it was difficult for an expert to assess directly knowledge on certain criteria. Therefore, we propose to conduct a thorough analysis of this entities intervening in the knowledge assessment. In this section we define the concepts of Knowledge, Actor, Support and Criteria extracted from the database (figure 1) by reusing ontological categories defined in the literature.

A. *Domain analysis*

1) *Knowledge*

The notion of knowledge identified in dictionaries means the ability of an individual (according to his learning faculty and memory) to analyze and understand information in order to assimilate and generate an interpretation and an own representation (tacit or explicit) with the intention to act in a given context. Each individual has his own knowledge. Each one represents the world in its own way and this representation determines how it addresses the problems (in accordance with the interests of time, mood, etc.). As such, we consider knowledge as a set of beliefs held by an individual (or several). In reference to the Belief-Desire-Intention paradigm, beliefs reflect the knowledge that can have an individual on the universe to which he/she belongs.

This acceptation of knowledge is closer to the one of the notion of *Computed Belief* which is defined in the COM ontology [12]. The principle is that an intentional agent have a *Mental State* (e.g., a *Belief*) about a *Mental Object* (respectively, a *Computed Belief*) at a certain time. Our notion of the knowledge rejoins that of *Proposition* defined in the I&DA ontology [29] or that of *Description* of the D&S ontology [35]. A *Proposition/Description* represents a mean for an individual to describe situations that he/she considers as existing in the world. In particular, a *Proposition* may correspond to the content of a document (this is important for the follow).

To lead a more effective analysis, we require to characterize and locate knowledge. Thus, in the context of activities within the car company, we mainly distinguish three different types of knowledge needed to control processes and which can be sensitive and crucial to the organization in question. They are:

- knowledge about the development and the adaptation of material resources necessary to lead the activity (e.g., knowledge about the adaptation of a chemical model, knowledge about the development of a simulation tool able to predict the rate of diesel dilution in oil)
- knowledge necessary to lighten some technical constraints of the activity: it is used indirectly in the activity and produced outside of the project;
- knowledge produced or used during the activity (e.g., knowledge about the improvement of strategies related to the supervisor).

We can differentiate more specifically two main categories of knowledge necessary for the control of sensitive process:

- the knowledge *produced during an activity* may be produced either intentionally or not. They may therefore be a desired outcome or result of a "side effect" of the activity, not predictable a priori. They were produced (in the sense that they are new knowledge) or processed during this activity (e.g., an updating of knowledge).
- the knowledge *used during an activity* (directly or indirectly such as the knowledge necessary to lighten some technical constraints of the activity, the knowledge about the development and the adaptation of material resources necessary to lead the activity) provide a help for an agent (or several) to carry out an action (to reach a goal).

This distinction allow us to precise our definitional framework by assimilating knowledge *produced during an activity* and knowledge *used during an activity* respectively to *artificial entities* and *functional entities* in the broadest sense of the definitions given in a recent study [37].

Finally, the UML model (see figure 1) distinguishes two classes of knowledge: the *Tacit-Knowledge* class (linked to the *Actor* class) and the *Explicit-Knowledge* class. We show in the following sections that we are doing the distinction between knowledge held by an individual (or several) (the *Tacit-Knowledge* class) (§2) and those inscribed on a support (the *Explicit-Knowledge* class) (§3). If we go back to the UML definition of the *Knowledge* class, a knowledge is defined by eight particular attributes: the criteria of scarcity, transferability, imitability, accessibility, complexity, validity, substitutability and the cost and time of knowledge production. In the section 4, we try to clarify the nature of these criteria necessary to classify the potential crucial knowledge.

2) Actors

When we are interested in the knowledge assessment, we consider the organization where knowledge is mobilized and used by different actors.

According to [26], the fact to know is similar to fact to be likely to act. A *knowledge* is therefore "actionable" and "to be likely to act" joined the concept of ability (or faculty) to perform an action. From a consensual point of view in AI, the notion of ability implies that knowledge is in an “ideal” level (it's a "private" experience) and belongs to mental world proper to an individual. It therefore does not coincide with any of the actions carried out:

As a potential (or ability, talent), the ability exists independently of the action to which it relates and whether that action succeeds or fails, then regardless of whether the result exists or not. [37]

We talk more specifically about competence, ability or talent of an individual. The knowledge is therefore "owned" by an individual (or several) giving him/her the ability to perform (and to repeat) an action (to reach a goal).

This faculty to perform an action is embodied in an entity defined in [39], the *Agentive*. An *Agentive* could be a human being, a robot, a knowledge-based system or an organization. He/she acts with the intent to achieve a goal and implements the appropriate means to achieve his goals. An *Agentive* plays the role of *Agent* (in the sense of [32]) during the *Action* in which it participates (according to the relationship of participation of the DOLCE ontology [28]). This means that an individual intern or extern to the organization (e.g., the French car company) could be both a decision maker and may be also a knowledge provider and/or a project manager. The notions of actor presented in the section 2 (*Knowledge provider, Project manager, Decision maker, Knowledge manager* and *Analyst*) are therefore specialized roles of *Agent* in the context of the process of knowledge evaluation.

3) Support

In our analysis, the knowledge results from the interpretation (the sense) given to any entity (an object, a process) by an individual or a social group (i.e., a community of intentional agents) in an organization. This knowledge is either owned by an individual or a social group (in the form of a mental inscription), or included on a support (e.g. a sheet of paper, an audio-visual document, a CD, the computer’s memories, etc.). The organizations have the “capacity” to give a status to certain objects (for example, a piece of paper can acquire the status of a bill because members of the organization recognized as such) [34].

According to the theory of the support of [44], a *Support* is a physical object having a semiotic inscription of knowledge (e.g. a text manuscript or printed materialized by some ink and formatted) intelligible for a cognitive agent (e.g. a human being, a software, a robot, etc.). This implies that the agent have the competences for interpreting the form perceived and give it meaning. This also implies that this form has been apprehended internally by the agent in the form of a mental inscription.

Therefore, the entity which makes sense is neither the document nor the object but an mental inscription resulting from the perception of the document by an agent. This can reflect the fact that objects that are not documents (which were not intentionally created as such) are not intrinsically sense but that agents can make meaningful perception of these objects. This can also reflect the fact that the nature of these objects can be any and it can include practices or temporal objects (like *Perdurants* in the meaning of DOLCE: which “happens in time” like processus, events, actions, etc).

For our needs, we restrict to the supports specially designed by the human to vehiculate and communicate meaning (database, knowledge base system). In other words, we only take into account neither natural objects nor artificial objects communicating accidentally sense (for example, the location of the moss on tree trunks informs on the direction of the wind in a geographical area).

4) Criteria of knowledge vulnerability

The UML definition of the *Knowledge* class is defined by a family of criteria (scarcity, transferability, imitability, accessibility, complexity, validity, substitutability and the cost and time of knowledge production) whose aims to influence the opinion of the decision makers about the cruciality of knowledge. However, within the meaning of the DOLCE ontology, these criteria could not be defined as properties (specifying the concept *Quality* defined in DOLCE) of the concept of knowledge because it is obvious that they are not “inherent” to a knowledge. This modeling choices have sense (i.e., they depend on the knowledge to which they assign) only for computing the contribution degree of knowledge in the studied context. This is confirmed by the definitions of the very heterogeneous criteria reminded in the section 2.3 of this paper: the *Scarcity* of the knowledge in the context of the knowledge evaluation amounts to assess the number of persons owning a certain knowledge, the *Cost and time of knowledge production* criterion represents the evaluated number of persons and the evaluated period needed to create the knowledge, and so on.

This definitional framework is that of a meta-level because it reflects the idea that a decision maker has about a given knowledge. In this consideration, we are interested in the work done by A. Gangemi in a technical report [16] about the ontology evaluation design pattern. In our turn, we could consider a quality-oriented knowledge description in which we would define parameters for the quality of knowledge (e.g., scarcity, complexity). However, for the same reasons evocated above, we cannot propose definitions of these criteria at a meta level.

A fact is that each definition of criteria assumes an action of knowledge assessment on certain criteria leading to transform the knowledge state of a decision maker. We assimilate the knowledge assessment on the different criteria to a reasoning which is decomposed into several sub-reasonings (for example, *Assessing knowledge scarcity* consisting in the assessment of the number of person who own the knowledge; *Assessing knowledge accessibility* consisting in the assessment of the time needed to access to the knowledge; and so on). This reasoning have for data a given knowledge to assess (for example, knowledge relative to material of filter support).

B. *Conceptual model*

On the basis of this previous analysis, we propose a modeling framework which consists in reusing the ontological resources defined by the team of G. Kassel in the MIS laboratory, extending the DOLCE ontology. These resources are available at the URL: http://www.laria.u-picardie.fr/IC/site/spip.php?article53. More precisely, our modelling framework exploits:

- the core ontology of *Actions*;
- the core ontology of *Participant roles* (also called "casuals roles" or "thematic roles" in the literature), which cover the domain of the "modes of participation" of the entities intervening in the evaluation of the crucial knowledge;
- the core ontology I&DA, which cover the domain of semiotics, initially built to classify documents by their contents.

By admitting that all knowledge is knowledge about "something", about an "object", we can schematically distinguish between two categories of knowledge, depending on the nature of the objects (physical or mental) with which it deals:

- *practical* knowledge (i.e. know-how “to act”) deals with *physical* objects and enables action in the real world (e.g. banging in a nail, riding a bicycle)
- *theoretical* knowledge (i.e. know-how “to think”) deals with *theoretical* objects (mental objects) and enables action in the mental world (e.g. calculating, deciding).

According to this definition and assimilating *Actions* to transformations of a world (entities), the core ontology of *Actions* divides actions into *Doings* (*Physical Actions*), which are actions on the physical world, and *Non-Physical Actions*, which aim at transforming the agent reasoner's mental world. It means that the modification does not concern the real world but the representation that the reasoner makes of the real world. Among the *Non-Physical Actions*, we distinguish the *Conceptual Human Actions* which transform *Conceptualizations* of a *Human agent* (e.g., *Assessing a hypothesis*, *Diagnosing a car’s breakdown*).

A *Conceptualization* is defined in the core ontology of I&DA. It is a mean by which agents can reason about a world. They are expressed in the form of *Expressions* and are physically realized by the *Inscriptions.* An *Inscription is* a knowledge form (e.g., printed texts, pictures) materialized by a substance (e.g., some ink, an electronic field) and inscribed on a Support, i.e. a material object (e.g., paper, hard disk, ambient air in the case where a text is read). An *Expression* is a non-physical knowledge form expressed in a communication code and for which an agent assigns some meaning. Among *Conceptualizations*, a functional distinction is made between *Propositions*, which are descriptions of situations, and *Concepts*, which allow for classifying entities in a world.

We define the action of evaluation of the crucial knowledge as a *Conceptual Human Action* which is an evaluation bearing on *knowledge produced during an activity* and knowledge *used during an activity*, and having for agent a *Decision maker*. This action of knowledge assessment is decomposed into several sub-actions (*Assessing knowledge scarcity*; *Assessing knowledge accessibility*, and so on) consisting in the evaluation of crucial knowledge on different criteria leading to transform the knowledge state of a decision maker.

The action of evaluation of crucial knowledge has for specific data a given knowledge to assess (for example, knowledge relative to material of filter support) which is a *Proposition* in accordance with the principle of modeling of I&DA. The ways of participation to an action are defined in the core ontology of *Participant roles* by introducing specialized relations of participation in the sense of DOLCE: only *Endurants* (i.e., an entity “enduring in time” as a pen, a car company, some water, human rights) participate in *Perdurants* (an entity which “happens in time” as the Olympics games, your reading of this article) and, furthermore, any *Endurant* participates necessarily to a *Perdurant*. For example, a *knowledge to assess* is a *Knowledge used during an activity* and an *Assessing Data*. The roles of *Assessing Data* and *Assessing Result* specialize classes of the *Data* and *Result* roles (figure 4).

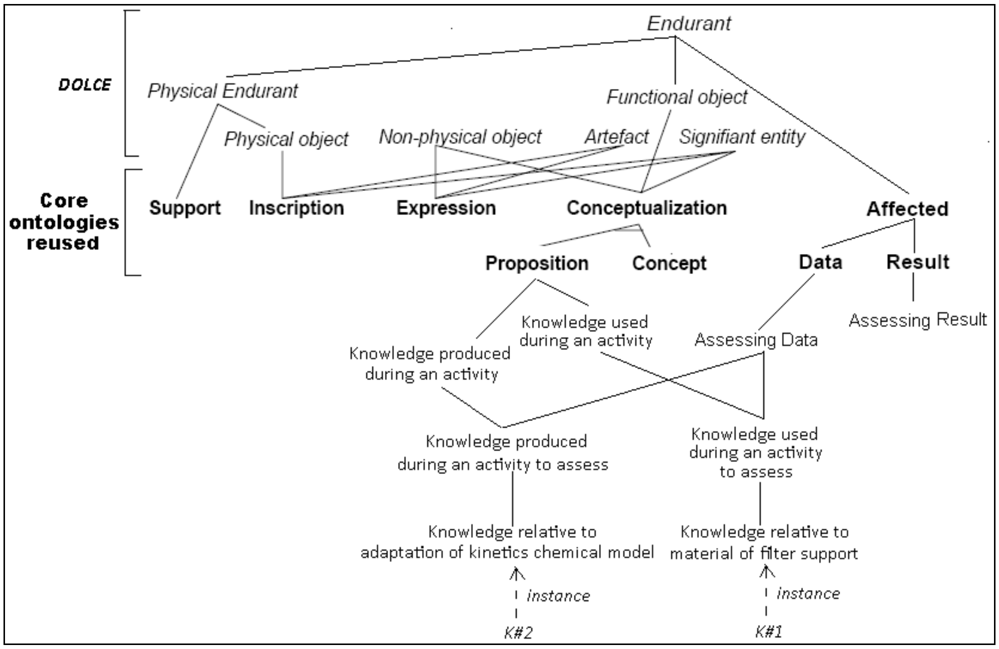

DOLCE
Core ontologies reused
Endurant
Physical Endurant
Physical object
Non-physical object
Functional object
Artefact
Signifiant entity
Support
Inscription
Expression
Conceptualization
Affected
Proposition
Concept
Data
Result
Knowledge used during an activity
Knowledge produced during an activity
Assessing Data
Assessing Result
Knowledge produced during an activity to assess
Knowledge used during an activity to assess
Knowledge relative to adaptation of kinetics chemical model
Knowledge relative to material of filter support
instance
instance
K#2
K#1

Figure 4. An excerpt of the ontology relative to the evaluation of crucial knowledge

The notion of *Agent* akins to a way for an *Endurant* to participate temporally in an *Action*. We specialize the role of *Agent* defined in the core ontology of *Participant Roles* to define the different contextual roles played by the members of the organization: the *Knowledge provider*, the *Project manager*, the *Decision maker*, the *Knowledge manager* and the *Analyst* (figure 5). For example, a *Decision maker* plays the role of agent in the action of *Making a decision*. It is important to mention that the same person may have different roles. For instance, a *Decision maker* may be also a *Knowledge provider* and/or a *Project manager*.

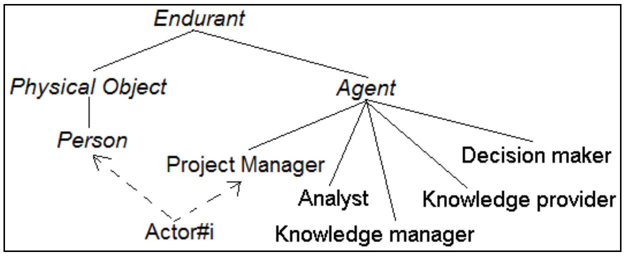


Figure 5. The sub-ontology of the persons concerned by the knowledge identification decision process; Actor#i *have for type* Person and *plays the role of* ProjectManager

V. Conclusion and future work

In this paper, we have done the first phase of the construction of an ontology covering the domain of the crucial knowledge identification. We have exploited the ontological categories existing in the literature to precise the definitions of the UML classes by proposing a coherent modeling framework linking the notions of Knowledge, Actor, Support and Criteria. In particular, we expressed that knowledge produced or used during an activity in the context of the car company is a *Proposition* participating as data and result in the evaluation of the cruciality of knowledge which is realized by a decision maker - an *Agent*.

Our work is currently continued to formalize the knowledge that we have defined in this article, the relations and the associated semantics. We aim to integrate into the system K-DSS the ontology of the domain of the crucial knowledge assessment to automatically reasoning from only a sample of crucial knowledge.

For example, assessing a knowledge *K* based on the complexity criteria is to study the number and the degree of dependency between the knowledge needed to maintain K. If the complexity of K is important, then K requires knowledge of at least four other knowledge, i.e., the different expertise or domain of knowledge of different businesses, used by an actor in a given activity. This process of evaluation on the complexity criteria is based on the tacit knowledge of decision makers. The idea is to make explicit such dependences (in the form of properties or relations in the ontology) between the crucial knowledge to enable the K-DSS system to automatically deduct the cruciality of a knowledge *Ki* knowing the cruciality of the knowledge *Kij*.

Finally, in order to optimize the collaboration between the system K-DSS and the end-users, and to use the system remotely, we will propose new features based on the principles of the technologies related to the Semantic Web (ontologies, reasoning, Web services, etc.). We will use the ontology editor Protege-2000, developed at Stanford University [30], which benefits of the development of “plug-in” for the languages RDF, DAML + OIL and OWL to specify an ontology in different languages on the Semantic Web. In terms of implementation, the plug-in JessTab integrated into Protege-2000, allows to introduce the knowledge stored by Protege-2000 in a database of facts for the application of rules by the inference engine Jess, to the instances of the ontology and the ontology itself (meta-reasoning).

REFERENCES

[1] J. Arpirez, O. Cororcho, M. Fernandez-Lopez, and A. Gómez-Pérez, WebODE in a nutshell, AI Magazine, 24(3), pp 37-48, 2003.

[2] N. Aussenac, Conception d'une méthodologie et d'un outil d'acquisition des connaissances expertes. Thèse de Doctorat en informatique, Université P. Sabatier, Toulouse, Octobre 1989.

[3] N. Aussenac-Gilles, B. Biébow and S. Szulman, D'une méthode à un guide pratique de modélisation de connaissances à partir de textes. In F. Rousselot, éditeur, Actes des 5e rencontres Terminologie et IA (TIA 2003), pages 41–53, Avril 2003.

[4] B. Bachimont, 2004. Art et sciences du numérique : ingénierie des connaissances et critique de la raison computationnelle. Mémoire d'habilitation à diriger des recherches, Université Technologique de Compiègne.

[5] S. Bechhofer, I. Horrocks, C. Goble, and R. Stevens, OilEd: a Reasonable Ontology Editor for the Semantic Web, In Proceedings of the Joint German/Austrian Conference on Artificial Intelligence (KI'2001), volume 2174, pp 396–408, Springer-Verlag LNAI, 2001.

[6] E. Bottazzi, and R. Ferrario, "Preliminaries to a DOLCE Ontology of Organizations", In International Journal of Business Process Integration and Management, Special Issue on Vocabularies, Ontologies and Business Rules for Enterprise Modeling. C. Atkinson, E. Kendall, G. Wagner, G. Guizzardi, M. Spies (Eds.), 2008.

[7] J. Breuker and B. Wielinga. KADS : Structured Knowledge Acquisition for Expert Systems. In Actes des 5èmes journées internationales "Les systèmes experts et leurs applications", Avignon, France, 1985.

[8] S. Bruaux S. and I. Saad, Improving semantic in the decision support system K-DSS. In Proceedings of the International Conference on Information, Process, and Knowledge Management. IEEE Computer Society Press p.p. 66-71, Cancun (Mexico), 1-7 February 2009.

[9] S. Bruaux, G. Kassel, and G. Morel, "A clarification of the ontological status of knowledge roles", In Proceedings of the Workshop on Advances in Conceptual Knowledge Engineering, co-located with the 18th International Conference on Database and Expert Systems Applications, Germany, Septembers 2007.

[10] T. H. Davenport, and L. Prusak, "Working Knowledge: How Organisations Manage What They Know" *Harvard Business School Press*, Boston, MA, 1998.

[11] R. Dieng, O. Corby, A. Giboin, and M. Rybière, Methods and tools for corporate knowledge management, Rapport technique, INRIA, projet ACACIA, Sofia, 1998.

[12] R. Ferrario and A. Oltramari, "Towards a Computational Ontology of Mind", Formal Ontology in Information Systems, Proceedings of the International Conference FOIS 2004, A.C. Varzi, and L. Vieu (Eds.), IOS Press Amsterdam, November 2004, pp. 287-297.

[13] M. Fernandez, A. Gomez-Perez, and N. Juristo, METHONTOLOGY: From Ontological Art Towards Ontological Engineering. AAAI-97 Spring Symposium on OntologicalEngineering, Stanford University, March 24-26th, 1997.

[14] J.Y. Fortier and G. Kassel, "Managing Knowledge at the Information Level: an Ontological Approach", In Proceedings of the ECAI'2004 Workshop on Knowledge Management and Organizational Memories, Valencia (Spain), August 2004, pp. 39-45.

[15] Fürst F., 2006. L'opérationalisation des ontologies : une méthodologie et son application au modèle des Graphes Conceptuels. In Journal électronique d'intelligence artificielle, Vol. 5, number 38, 2006.

[16] A. Gangemi and P. Mika, "Understanding the Semantic Web through Descriptions and Situations", Proceedings of the International Conference on Ontologies, Databases and Applications of Semantics (ODBASE 2003), R. Meersman, and al. (Eds.), Catania (Italy), November 2003.

[17] T.-R. Grüber. Towards Principles for the Design of Ontologies Used for Knowledge Sharing. In Nicola Guarino et Roberto Poly (eds.), editor, Proceedings of the International Workshop on Formal Ontologies, Padova, Italy, 1993. Kluwer Academic Publishers.

[18] M. Grundstein, C. Rosenthal-Sabroux and A. Pachulski, Reinforcing Decision Aid by Capitalizing on Company's Knowledge, European Journal of Operational Research, 145, pp. 256-272, 2003.

[19] M. Gruninger and M. S. Fox, Methodology for the design and evaluation of ontologies. In Proceedings of the Workshop on Basic Ontological Issues on Knowledge Sharing, IJCAI'95, 1995.

[20] N. Guarino and C. Welty, A formal ontology of properties. In R. Dieng et O. Corby, éditeurs, 12th International Conference in Knowledge Engineering and Knowledge Management (EKAW'00), pages 97–112. Springer Verlag, 2000.

[21] N. Guarino and P. Giaretta, Ontologies and knowledge bases: towards a terminological clarification. In N. Mars (Ed.), Towards very large knowledge bases (p. 25 - 32). Amsterdam IOS Press, 1995, http:// ontology.ip.rm.cnr.it/Papers/KBKS95.pdf.

[22] J. Kahan, M. Koivunen, E. Prud'Hommeaux, and R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, In Proceedings of the 10th International World Wide Web Conference, pp 623-632, 2001.

[23] G. Kassel. Le projet AIDE : une contribution aux systèmes experts de seconde génération. Mémoire d'Habilitation à Diriger des Recherches, Dauphine, 1995.

[24] G. Kassel, P. Lando, A. Lapujade, and F. Fürst, "Des Artefacts aux Programmes", In Proceedings of the 1ères Journées Francophones sur les Ontologies : JFO 2007, Sousse (Tunisia), 18-20 October 2007, pp. 281-300.

[25] G. Kassel, Integration of the dolce top-level ontology into the ontospec methodology, 2005. LaRIA Research Report 2005-08, Université de Picardie Jules Verne. Available at <http ://hal.ccsd.cnrs.fr/ccsd-00012203>.

[26] D. Kayser, La représentation des connaissances, Collection informatique, Hermès, Paris, 1997.

[27] DB. Lena and RV. Guha, Building Large Knowledge-based Systems: Representation and Inference in the Cyc Project. Addison-Wesley, Boston, Massachusetts, 1990.

[28] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari, and L. Schneider, "The WonderWeb Library of Foundational Ontologies and the DOLCE ontology", WonderWeb Deliverable D18, Final report (vr 1.0, 31-12-2003), 2003.

[29] F. N. Noy and D.L. McGuinness, Ontology Development 101: A Guide to Creating Your First Ontology, Technical Report SMI-2001-0880, Stanford Medical Informatics,Stanford University, Stanford, CA , USA.

[30] N. Noy, R. Fergerson, and M. Musen, The knowledge model of Protégé-2000: Combining interoperability and flexibility. In R. Dieng and O. Corby, editors, Proceedings of the 12th International Conference on Knowledge Engineering and Knowledge Management: Methods, Models, and Tools (EKAW 2000), volume 1937 of Lecture Notes in Artificial Intelligence (LNAI), pages 17–32, Juan-les-Pins, France, 2000. Springer.

[31] R. Neches, R. Fikes, T. Finin, T. Gruber, R. Patil, T. Senator and W.-R. Swartout, Enabling technologies for knowledge sharing. In AI Magazine, 12(3):36--56, Fall 1991.

[32] R. Volz, D. Oberle, S. Staab, and B. Motik: KAON SERVER - A Semantic Web Management System. WWW (Alternate Paper Tracks) 2003

[33] I. Saad, Une contribution méthodologique pour l'aide à l'identification et l'évaluation des connaissances nécessitant une opération de capitalisation. Ph.D. thesis, Université Paris-Dauphine, Paris, France, 2005.

[34] I. Saad and S. Chakhar, A decision support system for identifying crucial knowledge requiring capitalizing operation, European Journal of Operational Research (EJOR),Volume 195, n° 3, pp. 889-904, June 2009.

[35] G. Schreiber, B.J. Wielinga, H. Akkermans, W. Van de Velde, and A. Anjewierden. CML: The CommonKADS Conceptual Modelling Language. In Proceedings of the 8th European Knowledge Acquisition Workshop: EKAW'94, Springer-Verlag, 1994, p. 283-300.

[36] Sowa J., Conceptual structures : information processing in mind and machine, Addison-Wesley, 1984.

[37] S. Staab, H.-P Schnurr, R. Studer, and Y. Sure, Knowledge processes and ontologies. IEEE Intelligent Systems, Special Issue on Knowledge Management, 16(1). Staab S, Schnurr HP, Studer R, Sure Y (2001) Knowledge Processes and Ontologies. IEEE Intelligent Systems 16 (1): 26-34.

[38] L. Steel, Corporate knowledge management, Proceedings of the International Symposium on the Management of industrial and corporate knowledge (ICMICK'93), Compiègne, octobre 1993, pp.9-30.

[39] Y. Sure, S. Staab, and J. Angele, OntoEdit: Guiding ontology development by methodology and inferencing. In Proceedings of the International Conference on Ontologies, Databases and Applications of SEmantics ODBASE 2002, University of California, Irvine, USA, 2002.

[40] B. Swartout, R. Patil, K. Knight and T. Russ, Towards Distributed Use of Large-Scale Ontologies. Spring Symposium Series on Ontological Engineering, pp.138-148.

[41] B. Tseng and C. Huang, "Capitalizing on Knowledge: A Novel Approach to Crucial Knowledge Determination," IEEE Transactions on Systems, Man, and Cybernetics Part A: Systems and Humans, Volume 35, Issue 6, 919-931, 2005.

[42] M. Uschold and M. King, "Towards a methodology for building ontologies" In Workshop on Basic Ontological Issues in Knowledge Sharing: International Joint Conference on Artificial Intelligence. (Also available as AIAI-TR-183 from AIAI, The University of Edinburgh.), 1995.

[43] M. Uschold and M. Grüninger. Ontologies: Principles, Methods and Applications. In Journal of Knowledge Engineering Review, 11(2), 1996.

[44] B. Wielinga, J. Benjamin, W. Jansweijer, G. Schreiber, E. Meis, G. Willumsen, J. Eggen, P. Gobinet, N. Modiano, A. Bemaras, I. Laresgoiti, and F. Persson, Principles and Guidelines for Domain Ontology Library Design, v. 2, ESPRIT Project 8145 KACTUS, deliverable DO5a.2, 1996.

Towards Semantic Exchange of Clinical Documents

Juha Puustjärvi
Helsinki University of Technology
Espoo, Finland
juha.puustjärvi@tkk.fi

Leena Puustjärvi
The Pharmacy of Kaivopuisto
Helsinki, Finland
leena.puustjärvi@kolumbus.fi

Abstract— **Nowadays healthcare institutions have major problems with accessing and maintaining the large amounts of data that are continuously being generated. In addition, system interoperation is of prime importance as much of the patients' relevant information may be historic, and may have been gathered over many encounters with healthcare providers in different locations using heterogeneous healthcare information systems. In order to promote system interoperation several organizations in the healthcare sector have produced standards and representation forms using XML. However, the introduction of these XML-based technologies is not enough to provide a means to interpret the semantics of the exchanged messages. As a result, extending systems by new parties as well as introducing new message types is inconvenient. Replacing existing hard-coded medical information systems by open healthcare information systems that support semantic interoperation, are extensible, and maintainable is a challenging research problem. In this article we have restricted ourselves on this problem. In particular, we described our work on using RDF in exchanged clinical documents. Such documents themselves describe their semantics, and so they are in a machine understandable form. Hence RDF-based messaging represents an open, easily maintainable and extensible way for developing interoperable open systems.**

Keywords*- e-health; open healthcare systems; semantic interoperability; ontologies.*

I. INTRODUCTION

Information and communication technology has not only changed the way that clinical documents are stored and generated across and within healthcare organizations but it has also increased the efficiency and cost-effectiveness of healthcare organizations.

In particular, during the past few years the technology developed for interoperable autonomous systems has significantly changed. In particular, XML is rapidly becoming the key standard for data representation and transportation. However, the existing medical information systems that have been built during the past decades are based on proprietary solutions, developed in piecemeal way, and tightly coupled through ad hoc means [1]. These systems have many duplicated functions, and they are monolithic, non-extensible and non-interoperable [2, 3, 4, 5]. Such systems are commonly called *stovepipe systems* [6] as their components are hard-coded to only work together.

How to replace the stovepipe systems by the open healthcare information systems that support semantic interoperability, are extensible and maintainable is a challenging problem for the healthcare sector.

In our research we have focused on this problem. In particular, our focus is semantic exchange of pharmaceutical information between medical information systems. By semantic exchange we refer to the ability that the communicating parties can unambiguously (based on medicinal ontologies) interpret the exchanged messages [7]. By a medical information system [8] we refer to any system that processes medical data.

The starting point for our work has been the goal to develop an experimental infrastructure for exchanging pharmaceutical information, which satisfies the following goals: (i) The system supports semantic interoperability. (ii) Communicating information systems can independently introduce new message types. (iii) The system is open in the sense that new participating medical information systems can be easily introduced.

Whether these goals can be achieved by utilizing Semantic Web-technologies [7] is the main topic of this article. The article extends the work presented in [1].

The rest of the article is organized as follows. First, in Section II, we characterize open systems as they comprise the cornerstone of our approach. Then, in Section III, we give a short overview of the state of the art with respect to exchanging medical information. We first consider how semantic interoperability is achieved in the CDA (Clinical Data Architecture) [9] by hard-coding the semantics of the messages in communicating systems. Then, we consider the use of XML-based messaging in electronic prescription systems, and illustrate why XML-based messaging [10] requires hard-coding. In Section IV, we consider RDF-based messaging, i.e., message exchange where the messages include RDF-statements [11]. In such messages the semantics of the message is available from external sources, and so there is no need for hard-coding. The architectural and technical aspects of RDF-based messaging are considered in Section V. Finally, Section VI concludes the article by discussing the advantages and disadvantages of the deployment of the RFF-based technology in exchanging clinical documents.

II. OPEN SYSTEMS

O*pen systems* are computer systems that provide some combination of interoperability, portability, and open software standards [12]. In this article we consider open systems from interoperability point of view. By interoperability we refer to the ability of diverse systems and organizations to work together.

A. Semantic interoperability

Shared understanding of the exchanged messages can be achieved by *semantic interoperability*, which means that after data were transmitted from a sender system to a receiver, all implications made by one party had to hold and be provable by the other [12].

There are two thoroughly different approaches for achieving semantic interoperability: hard-coding and semantic messaging.

By *hard-coding* we refer to the software development practice of embedding the semantics input-messages into the application program, instead of obtaining the semantics from external sources. Hard-coding is proven to be a valuable and powerful way for exchanging structured and persistent business documents. However, if we use hard-coding in the case of non- persistent documents and non-static environments we will encounter problems in deploying new document types and extending the system by new participants.

By *semantic messaging* we refer to the practice of including the semantics of the exchanged document in a machine understandable form in the messages. Exchanging semantic messages represents an open, easily maintainable and extensible way for developing interoperable open systems.

B. Autonomy and heterogeneity in open systems

The emerging open information systems are co-operative where autonomous and heterogeneous components enable the components collectively to provide solutions. This requires that the information systems have components that cross organizational boundaries, and in this sense are open. In open systems the components are autonomous and heterogeneous, and the configuration of the whole system can change dynamically.

Fundamentally components´ autonomy means that they function under their own control. The reason for this is that the components reflect the autonomy of the organization interests that they represent. In addition there may be technical reasons for the autonomy, e.g., as a result of a hardware failure or error in a software.

In open systems heterogeneity can arise in a variety of formats, e.g., in networking protocols, in encoding information, and in used data models. Heterogeneity may also arise at semantic levels, e.g., the same concept is used for different meanings, or two different concepts are used for the same meaning. The reason for heterogeneity is historical: the components may have arisen out of legacy systems that are initially developed for local uses, but are eventually expanded to participate in open environments.

C. Document-centric Web services

SOA (Service Oriented Architecture) is an architectural design pattern that concerns with defining loosely-coupled relationships between producers and consumers [12]. It provides flexible methods for connecting information systems themselves as well as to other relevant systems. SOA relies on Web services as its fundamental design principle.

Technically Web services are self-describing modular applications that can be published, located and invoked across the Web. Once a service is deployed, other applications can invoke the deployed service.

There are two ways of using Web services: the RPC-centric view (Remote Procedure Call–centric) and the document-centric view. The *RPC-centric view* treats services as offering a set of methods to be invoked remotely while the *document–centric view* treats Web-services as exchanging documents with one another. Although in both approaches transmitted messages are XML-documents, there is a conceptual difference between these two views.

In the RPC-centric view the application determines what functionality the service will support, and the documents are only business documents on which the computation takes place. Instead the document-centric view considers documents as the main representation and purpose of the distributed computing: each component of the communicating system reads, produces, stores, and transmits documents. The documents to be processed determine the functionality of the service. Therefore, document centric view corresponds better with our goal of applying services in open environments. Furthermore, as the RDF-based messages are also represented by XML, the document-centric view suits well for semantic exchange of clinical documents.

III. EXCHANGING CLINICAL DATA

Health care systems are designed to meet the health care needs of target populations. The goals for health systems are good health, responsiveness to the expectations of the population, and fair financial contribution.

There are a wide variety of health care systems. In many countries health care system has evolved and has not been planned. Most of the health care systems that are developed during the past decades are closed. Typically they are proprietary and only serve one specific department within a healthcare institution [13]. Such standalone systems are developed by many different suppliers, and thus they are incompatible with one another. However, this is regrettable as system interoperation is crucial since much of the patients' relevant information may be historic, and may have been gathered over many encounters with healthcare providers in different locations using heterogeneous healthcare systems.

In order to improve interoperation, several organizations in the healthcare sector have produced standards and representation forms using XML. For example, patient records, blood analysis and electronic prescriptions [14, 15, 16, 17, 18] are typically represented as XML-documents. The introduction of these XML-based technologies alleviates the stovepipe problem but they are not enough to achieve semantic interoperability. Instead for achieving semantic interoperability it is necessary to provide standardized ways to describe the meanings of the exchanged XML-documents.

A. Interoperation in HL7 CDA

The Clinical Document Architecture (CDA) is an ANSI approved HL7 standard [19] for the exchange, integration, sharing, and retrieval of electronic health information that supports clinical practice and the management, delivery and evaluation of health services. "Health level seven" refers to the seventh (application) level of the International Organization for Standardization (ISO) seven-layer communications model for Open Systems Interconnections.

The HL7 Clinical Document Architecture (CDA) is a document markup standard that specifies the structure and semantics of clinical documents for the purpose of exchange. Release One (CDA R1), became an American National Standards Institute (ANSI)–approved HL7 Standard in 2000. Release Two (CDA R2) [20, 21], became an ANSI-approved HL7 Standard in 2005.

A CDA document is a defined and complete information object that can include text, images, sounds, and other multimedia content. CDA documents are encoded in Extensible Markup Language (XML), and they derive their meaning from the HL7 Reference Information Model (RIM) and use the HL7 Version 3 Data Types.

RIM is static object-oriented model in UML notation. It serves as the source from which all specialized HL7 version 3 information models are derived and from which all HL7 data ultimately receives its meaning.

HL7 is proven to be a valuable and powerful standard for a structured exchange of persistent clinical documents between different software systems. However, in the case of non persistent documents with CDA we encounter many problems.

The reason for this is that the semantics of the documents is bound to the shared HL7 Reference Information Model (RIM) [19]. The developers of the CDA-compliant systems are familiar with the RIM and use that information in developing CDA-compliant systems. That is, HL7 compliancy means that the knowledge of the relationship between the XML-elements in the received CDA document and the conceptual schema given in RIM is hard-coded in the systems receiving the messages. Therefore HL7 CDA compliant systems are able to understand each other as long as they exchange CDA-documents.

The semantics of the CDA-compliant message cannot be interpreted just based on the message and the conceptual schema given in RIM. Therefore introducing a new message-type (i.e., a CDA document) and corresponding extensions to RIM is a long lasting process requiring standardization and the modifications of the communicating software modules. As a result, applying HL7 standards to a new domain, (e.g., for pharmacy) is problematic. Therefore the solutions made in the HL7 CDA standard do not satisfy the goals of open, extensible healthcare information systems that support semantic interoperability

B. Electronic Prescriptions

Electronic prescription is the electronic transmission of prescriptions of pharmaceutical products from legally professionally qualified healthcare practitioners to registered pharmacies [22]. The scope of the prescribed products varies from country to country as permitted by government authorities or health insurance carriers.

The information in an electronic prescription includes for example, prescribed products, dosage, amount, frequency and the details of the prescriber. A simple prescription is presented in Fig. 1 in XML.

```
<prescription>
        <prescription_id>abc123</prescription_id>
        <patient>
                <name>John Smith </name>
                <id> 1465766677</id>
        </patient>
        <medicinal_product>Panadol</medicinal_product>
        <disease>fever</disease>
        <quantity>30</quantity>
        <dose>One tablet three times a day</dose>
        <physician>
                <name>Lisa Taylor </name>
                <id> 98765432</id>
        </physician>
</prescription>
```

Figure 1. A simplified prescription in XML.

C. The Semantics of Prescriptions

XML (Extensible Markup Language) is a set of rules for encoding documents electronically. XML's design goals emphasize simplicity, generality, and usability over the Internet.

Although XML-documents are commonly used for information exchange they do not provide any means of talking about the semantics (i.e., meaning) of data. For example there is no meaning associated with the nesting of the tags presented in the XML-coded prescription in Fig. 1. It is up to the applications that receive the XML-messages to interpret the nesting of the tags. Even if there is a conceptual schema or ontology [23, 24] having the modeling primitives having the same naming (e.g., class patient having attribute name) as the tags in the XML-message it is up to the application to interpret the nesting of tags. To illustrate this consider the statement:

"Physician Lisa Taylor cares for patient John Smith".

We can present this sentence by the following two nesting ways:

```
(1)   <patient name='John Smith'>
              <physician>Lisa Taylor</physician>
      </patient>

(2)    <physician name='Lisa Taylor'>
              <patient>John Smith</patient>
       </physician>
```

These formalizations include an opposite nesting although they represent the same information. Hence, there is no standard way of assigning meaning to tag nesting.

Therefore the semantics of the documents in the messages (e.g., the prescription) must be specified by binding it to a conceptual schema (ontology), e.g., to a conceptual schema presented in Fig. 2.

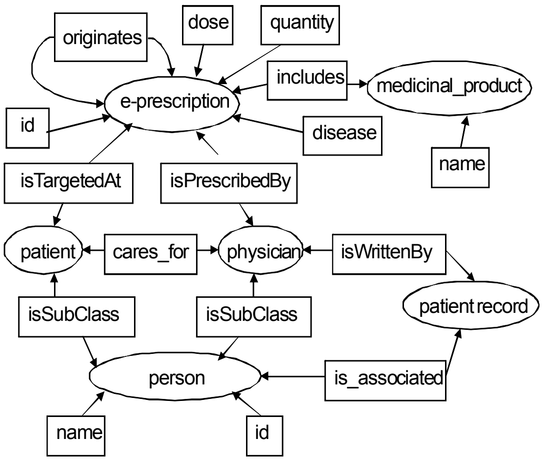


Figure 2. A medicinal ontology.

D. *Introducing New Message-Types*

In the case of hard-coded message exchange, the introduction of a new XML-message type requires that the syntax and the semantics of the message must be first standardized, and then the communicating systems' Web services can be updated by a new message type, i.e., the semantics of the messages can be hard-coded to the communicating applications.

In order to illustrate the problems of such hard-coded solutions, assume that the communicating medicinal systems do not only exchange electronic prescriptions but also renewed prescriptions. A renewed prescription deviates from other prescription in that it equals with the original prescription with respect to medicinal product but may deviate with respect to prescribing physician, quantity and dose. Such a renewed prescription of the prescription of Fig. 1 is presented in Fig. 3.

```
<prescription>
        <originates_id>abc123</originates_id>
        <patient>
                <name>John Smith </name>
                <id> 1465766677</id>
        </patient>
        <medicinal_product>Panadol</medicinal_product>
        <disease>fever</disease>
        <quantity>50</quantity>
        <dose>Two tablet three times a day</dose>
        <physician>
                <name>Paul Goodman </name>
                <id> 66765555</id>
        </physician>
</prescription>
```

Figure. 3. A renewed prescription presented in XML.

In order that the communicating medical information systems (e.g., electronic prescription writer, medical expert system, medical database system and electronic prescription holding store) would understand the syntax and semantics of the renewed prescription the structure of the XML-document should be standardized and its semantics should be specified by the conceptual schema. In addition, the semantics of the renewed prescription should be hard-coded in communicating information systems.

Another approach for deploying renewed prescriptions is that the renewed prescription message itself describes its semantics, and hence no message standardization process is needed. How this can be done by RDF is the topic of the next sections.

IV. RDF-Based Messaging of Medicinal Data

A. *RDF-Based Prescriptions*

The Resource Description Framework (RDF) is a language for representing information about resources in the World Wide Web. It is also a data model. Principally the RDF data model is not different from classic conceptual modeling approaches such as Entity-Relationship or Class diagrams, as it is based upon the idea of making statements about resources. It has come to be used as a general method for conceptual description or modeling of information that is implemented in web resources.

RDF provides a simple language in which to capture knowledge. It incorporates a number of well-known ideas from knowledge representation. RDF is built on top of the Web notion of a URI (Universal Resource Identifier). URIs need not be absolute in that they need not correspond to the name of any actual object to be accessed via any specific protocol.

RDF's modeling primitive is an object-attribute-value triple, which is called a statement [7]. For example, the preceding sentence "Physician Lisa Taylor cares for patient John Smith" is such a statement.

There are various ways in capturing knowledge with RDF, e.g., as natural language sentence as above, in a simple triple notation called N3, in RDF/XML serialization format, and by as a graph of the triples [7]. In Fig. 4, the prescription of Figure 1 is presented as a graph of triples, whereas in Fig. 5 it is presented in RDF/XML serialization format.

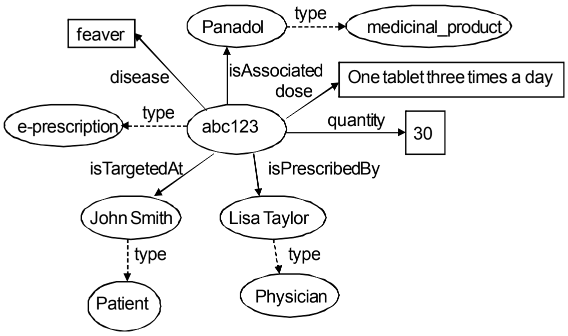


Figure 4. RDF-based prescription in a graphical form.

```
<rdf:RDF
    xmlns : rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns : xsd="http://www.w3.org/2001/XMLSchema#"
    xmlns : mo="http://www.lut.fi/ontologies/montology#"
    <rdf:Description rdf:about="abc123">
        <rdf:type rdf:resource="&mo;e-prescription"/>
            <mo : dose>One tablet three ti mes a day</mo : dose>
            <mo : quantity rdf:datatype="&xsd;integer">30</mo : quantity>
            <mo: includes>Panadol</mo: includes>
    </rdf : Description>
    <rdf:Description rdf:about="1465766677">
        <rdf:type rdf:resource="&mo;patient"/>
            <mo : name>John Smith</mo : name>
    </rdf : Description>
    <rdf:Description rdf:about="98765432">
    <rdf:type rdf:resource="&mo;physician"/>
            <mo : name>Lisa  Taylor</mo : name>
    </rdf : Description>
</rdf:RDF>
```

Figure 5. An electronic prescription in RDF-format.

B. *RDF-Schema and RDF-Typing*

RDF is domain-independent in that no assumptions about a particular domain of use are made. It is up to users to define their own domain specific terminology (vocabulary) by RDF Schema (RDFS) [7].

Our defined medicinal vocabulary (ontology) includes concepts patient, physician, e-prescription and patient record as well as their relationships. Basically it deviates from the ontology presented in Figure 2 in that it is presented by RDFS. Using the medicinal vocabulary we can state for example "Physician Lisa Taylor cares for patient John Smith" in a machine understandable way. Particularly by using the RDF-type element we tie the subject, predicate and the object of the statement "Physician Lisa Taylor cares for patient John Smith" to the RDF Schema. To illustrate this consider Fig. 6, which includes a subset of the ontology presented in Fig. 2 and the RDF-statement "Physician Lisa Taylor cares for patient John Smith".

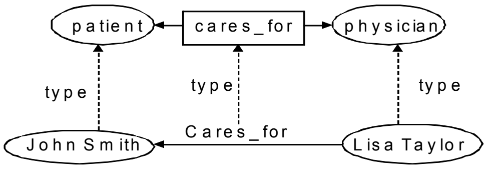


Figure 6. Typing an RDF-statement by RDFS.

V. The Architecture of the Communicating Medical Information Systems

A. *The Components of the Architecture*

In our used architecture medical information systems communicate through Web services by the SOAP -protocol. The semantic exchange of clinical documents is carried out by the SOAP-messages [12], which include RDF-statements. The components of the architecture are presented in Fig. 7.

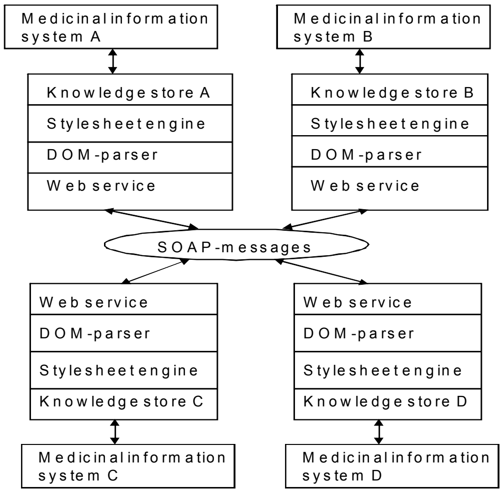


Figure 7. The components of the communicating systems

We next consider the components of the architecture from technology point of view.

B. *Web services*

As we have already stated the document-centric view of Web services suits for our purposes. The implementation of document-centric Web services of a prescription holding store is illustrated in Fig. 8. Here the Web service supports three kinds of requests: e-prescription requests, requests on patient's records and requests on patients' prescriptions. Each type of request is presented by specific document that is presented in RDF. However, the Web service does not support separate operations for these requests but rather a single operation, which just receives the documents and stores them in the Knowledge store. Further, processing the requests is the function of the Prescription management application.

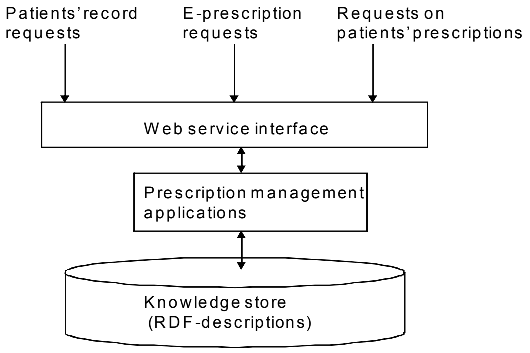

Patients' record requests
E-prescription requests
Requests on patients' prescriptions
Web service interface
Prescription management applications
Knowledge store (RDF-descriptions)

Figure 8. The structure of a Document-centric Web service.

A consequence of our used document-centric view is that we have to model the requests in the ontology of the Knowledge store; otherwise we could not store and retrieve the requests.. As the schema of the Knowledge store is specified by RDFS [7, 11], we have to model the requests also in RDFS. That is, we have RDFS class Request and its subclasses E-prescription request, Request on patient's record and Request on patients' prescription.

C. SOAP-messaging

SOAP was originally intended to provide networked computers with remote-procedure call services written in XML. It has since become a simple protocol for exchanging XML-messages over the Web.

A SOAP-message is comprised of a SOAP header, SOAP envelope and SOAP body. In particular, the SOAP body contains the application-specific message that the backend application will understand. As illustrated in Fig. 9, we incorporate our used RDF-formatted clinical documents in the SOAP body.

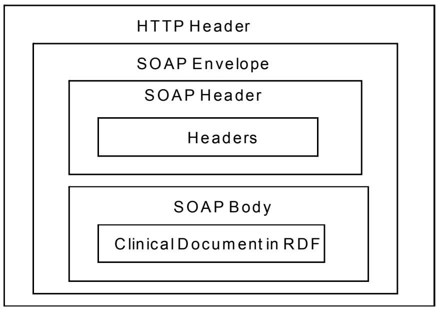


Figure 9. RDF-formatted clinical document in a SOAP-message.

An example of XML-coded SOAP-message which contains an RDF-formatted clinical document is presented in Fig. 10.

```
<SOAP-ENV: Envelope
  xmlns:SOAP-ENV=“http://schemas.xmlsoap.org/soap/envelope/”
  SOAP-ENV:encodingStyle=”http://schemas.xmlsoap.org/soap/encodig/”>
     <SOAP-ENV:Body>
        <clinical-document>
             <rdf:RDF
                xmlns : rdf=”http://www.w3.org/1999/02/22-rdf-syntax-ns#”
                xmlns : xsd=”http://www.w3.org/2001/XMLSchema#”
               xmlns : mo=“http://www.lut.fi/ontologies/montology#”
               <rdf:Description rdf:about=”abc123”>
                   <rdf:type rdf:resource=“&mo;e-prescription”/>
                    <mo : dose>One tablet three ti mes a day</mo:dose>
                    <mo : quantity rdf:datatype=”&xsd;integer”>30</mo:quantity>
                    <mo: includes>Panadol</mo: includes>
                 </rdf : Description>
                 <rdf:Description rdf:about=”1465766677”>
                     <rdf:type rdf:resource=“&mo;patient”/>
                     <mo : name>John Smith</mo:name>
                  </rdf : Description>
                  <rdf:Description rdf:about=”98765432”>
                     <rdf:type rdf:resource=“&mo;physician”/>
                     <mo : name>Lisa  Taylor</mo:name>
                  </rdf : Description>
             </rdf:RDF>
          </clinical-document>
       </SOAP-ENV: Body>
</SOAP-ENV: Envelope>
```

Figure 10. An RDF-formatted prescription in a SOAP-message

The RDF-coded clinical document of Fig. 10 is the prescription presented in Fig. 5. The namespaces "mo" specifies the used ontology. That is, the namespace "mo" refers to the URL where the ontology of Fig. 2 is stored in RDFS.

D. Processing exchanged clinical documents

In order that the medicinal information systems are able to handle the clinical documents of the SOAP-messages they have to use the DOM-parser and the Stylesheet engine. The DOM parser transforms input text (i.e., RDF-statements) into a tree, which is suitable for the Stylesheet engine to process. DOM (Document Object Model) [6] refers to a language-neutral data model and application programming interface (API) for programmatic access and manipulation of XML-coded data. Generally, parsing (also called syntactic analysis) is the process of analyzing a sequence of tokens to determine its grammatical structure with respect to a given formal grammar.

As illustrated in Fig. 11, the Stylesheet engine takes the RDF-document from the DOM-parser, loads it into a DOM source tree, and picks out the needed information by transforming the RDF-document with the instructions given in the style sheet.

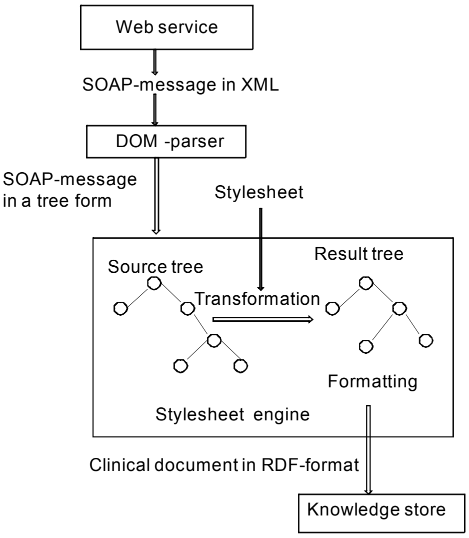


Figure 11. Transforming the representation formats.

In transforming the source tree the Stylesheet engine use XPath [6] expressions to reference portions of the tree and capture information to place it into the result tree. The result tree is then formatted, and the resulting RDF-document is stored in the Knowledge store.

E. Knowledge store

In our used architectural terminology Fig. 6 represents a overly simplified knowledge store in the sense that it includes and ontology represented in RDFS and one RDF-statement. In reality the knowledge stores have much wider ontology and thousands or millions of RDF-statements.

A salient feature of our used architecture is that the communicating medical information systems maintain their own knowledge store by picking out the interested knowledge from the messages they receive (i.e., RDF-statements) and then storing that knowledge to their own knowledge store.

To motivate this kind of message exchange strategy assume that the medical information system A sends a prescription to medical information systems B, C and D. These medical information systems may have different interests on the prescription.

For example, assuming that system B represents a pharmacy, so it is needs all the information in the prescription. On the other hand, assuming that system C represents government authorities, then it is obvious the system does not need information concerning the dose of the medicinal products; and assuming that system D represents health insurance authorities, then the system needs only the information of the patient and the prices of the medicinal products included in the prescription.

That is, each medicinal system has its own interest on the prescription, and they will only store in their knowledge store that part of the prescription. As illustrated in Fig. 11 the part on which a system has its interest is specified by the stylesheet it uses.

VI. CONCLUSIONS

Today healthcare institutions have major problems with accessing and maintaining the large amounts of data that are continuously being generated. At the same time the recent developments in the field of information technology have promised to bring improvements in the quality of managing and exchanging medicinal information. Also the technology developed for interoperable autonomous systems has significantly developed giving chances for implementing open healthcare information systems, which are easily extensible and maintainable.

In particular, the technology developed for interoperable autonomous systems has significantly changed. XML is rapidly becoming the key standard for data representation and transportation. However, the existing medical information systems that have been built during the past decades are based on proprietary solutions, developed in piecemeal way, and tightly coupled through ad hoc means

In this article, we have considered how to replace the hardcoded medical information systems by the open healthcare information systems that support semantic interoperability, and which are easily extensible and maintainable.

Semantic interoperability is the ability of computer systems to communicate information and have that information properly interpreted by the receiving system in the same sense as intended by the transmitting system. Semantic interoperability requires that any two systems will derive the same inferences from the same information.

The corner stone of our approach in achieving semantic interoperation is the medicinal ontology on which the communicating medical information systems have to commit in their mutual communication, i.e., the used medicinal ontology must be shared and consensual terminology as it is used for information sharing and exchange. It, however, does not suppose the introduction of a universal ontology for the healthcare sector. This situation is analogous with natural languages: a pharmacy, or any medicinal organization, may communicate in Finnish with medicinal authorities and in English with pharmaceutical companies. Just as there is no universal natural language, so there is no universal ontology.

A challenging situation for the health care organizations is also the introduction of new technologies. The introduction of semantic interoperation in healthcare sector is challenging as it incorporate semantic web technologies into many part of the work life cycle, including information production, presentation, analysis, archiving, reuse, annotation, searches and versioning. The introduction of these technologies also changes the daily duties of the many ICT-employees of the organization. Therefore the most challenging aspect will not be the technology but rather changing the mind-set of the ICT-employees and the training of the new technology.

The introduction of a new technology is also an investment. The investment on new Semantic Web-technology includes a variety of costs including software, hardware and training costs. Training the staff on Semantic Web-technology is a big investment, and hence many organizations like to cut on this cost as much as possible. However, the incorrect usage and implementation of a new technology, due to lack of proper training, might turn out to be more expensive in the long run.

REFERENCES

[1] J. Puustjärvi, and L. Puustjärvi. Semantic Exchange of Medicinal Data: a Way Towards Open Healthcare Systems. In the proc. of the Third International Conference on Digital Society (ICDS 2009), p.168-173.

[2] C. Liu, A. Long, Y. Li, K. Tsai, and H. Kuo, “Sharing patient care records over the World Wide Web”, International journal of Medical Informatics, 61, 2001, p. 189-205.

[3] R. Batenburg, and E. Van den Broek, “Pharmacy information systems: the experience and user satisfaction within a chain of Dutch pharmacies”, International Journal of Electronic Healthcare. Vol. 4, No.2 , 2008, p.119-131.

[4] K. Khoumbati, S. Shah, Y.K. Dwivedi, and M.H. Shah”, Evaluation of investment for enterprise application integration technology in healthcare organisations: a cost-benefit approach”, International Journal of Electronic Healthcare. Vol. 3, No.4, 2007, p.453-467.

[5] M.S. Raisinghani, and E. Young, “Personal health records: key adoption issues and implications for management”, International Journal of Electronic Healthcare. Vol. 4, No.1, 2008, p.67-77.

[6] M. Daconta, L. Obrst, and K. Smith. The semantic web. Indianapolis: John Wiley & Sons. 2003.

[7] G. Antoniou, & F. Harmelen. A semantic web primer. The MIT Press. 2004

[8] F. Jung, “XML-based prescription drug database helps pharmacists advise their customers”,
http://www.softwareag.com/xml/aplications/sanacorp.htm, 2005

[9] HL7 Overview. http://www.interfaceware.com/manual/hl7.

[10] E. Harold, and W. Scott Means W., XML in a Nutshell. O’Reilly & Associates, 2002.

[11] J., Davies, D. Fensel and F. Harmelen., Towards the semantic web: ontology driven knowledge management. West Sussex: John Wiley & Sons.2002.

[12] M. Singh, & M. Huhns. Service Oriented Computing: Semantics, Processes, Agents. John Wiley &Sons, Ltd. 2005.

[13] J. Puustjärvi, and L. Puustjärvi. Managing Medicinal Instructions. In the proc. of the International Conference on Health Informatics (HEALTHINF 2009), p.105-110.

[14] P. Woolman, XML for electronic clinical communication in Scotland. International journal of Medical Informatics, 64, 2001, p. 379-383.

[15] G. Stalidis, A. Prenza, N. Vlachos, S. Maglavera, D. Koutsouris, "Medical support system for continuation of care based on XML web technology", International journal of Medical Informatics, 64, 2001, p. 385-400.

[16] R. Keet, "Essential Characteristics of an Electronic Prescription Writer", Journal of Healthcare Information Management, vol 13, no 3.1999.

[17] J. Puustjärvi, and L. Puustjärvi "The challenges of electronic prescription systems based on semantic web technologies", In Proc. of the 1st European Conference on eHealth (ECEH'06). pages 251-261. 2006.

[18] J. Puustjärvi, and L. Puustjärvi, "Automating the coordination of electronic prescription processes", In Proc. of the 8th International Conference on e-Health Networking Applications and Services (HealthCom2006). p. 147-151. 2006.

[19] R.H. Dolin, L. Alschuler; C. Beerb, P.V. Biron, S. L. Boyer, D. Essin, E. Kimber, T. Lincoln, and J.E. Mattison. "The HL7 Clinical Document Architecture", J. Am Med Inform Assoc 2001:8(6), p.552-569.

[20] Robert H. Dolin, MD, Liora Alschuler, Sandy Boyer, BSP, Calvin Beebe, Fred M. Behlen, Paul V. Biron and Amnon Shabo (Shvo), HL7 Clinical Document Architecture, Release 2, http://www.jamia.org/cgi/content/abstract/13/1/30

[21] Gerdsen F, Müller S, Jablonski S, Prokosch HU. Standardized exchange of clinical documents--towards a shared care paradigm in glaucoma treatment.. Methods Inf Med. 2006;45(4):359-66.

[22] J. Puustjärvi, and L. Puustjärvi, "Developing an application integration strategy for electronic prescription system", In proc. of the International Workshop on Semantic Information Integration on Knowledge Discovery (SIIK2006), p. 253-262. 2006

[23] M. Gruber, Thomas R., "Toward principles for the design of ontologies used for knowledge sharing", Padua workshop on Formal Ontology, 1993.

[24] E. Mattocks, "Managing Medical Ontologies using OWL and an e-business Registry / Re-pository", http://www.idealliance.org/proceedings/xml04/papers/85/MMOOEBRR.html, 2005.

Remote Laboratory Access and Network simulation Tools for Students with Vision Impairment

Iain Murray
Department of Electrical & Computer Engineering
Curtin University of Technology
Perth Australia
i.murray@curtin.edu.au

Helen Armstrong
School of Information Systems
Curtin University of Technology
Perth Australia
h.armstrong@curtin.edu.au

Abstract—The delivery of laboratory exercises to students that are unable to attend in person due to physical disabilities is a significant issue. Both Netlab and Packet Tracer are inaccessible to many students who use assistive technology, particularly those with vision impairment. This paper presents the development of an accessible, cost effective, remote laboratory and describes the modification to laboratory sessions necessary for the blind to undertake Cisco Certified Network Associate (CCNA) laboratory sessions remotely and with full accessibility. Also discussed is the development of an accessible network simulator, iNetSim, illustrating possible methodologies that may be applied to make existing simulation packages accessible to those with severe vision impairment.[1]

Keywords-component; Networking laboratories, vision impaired, accessible eLearning

I. Introduction

The image of a vision impaired person holding a job on an IT Help Desk or in computer network administration seems a bit far-fetched: computers are vision driven and the vision impaired person would have difficulty seeing the user's screen to diagnose and fix any problem. Quite to the contrary, vision impaired people are ideal for this job role. Whilst computers have cables, plugs and plenty of ports, once they are installed computer networks need little physical attention. The attention they do require is of a logical nature, establishing and maintaining connections and access to the data required by the users. Users on business networks constantly need assistance from computer network professionals and this usually takes place via the IT Help Desk. Vision impaired people may have mobility problems and require extensive orientation and mobility training for each locale that they are required to work in. However technical support jobs require logical knowledge and skills, not physical mobility, making IT Help Desk positions ideal for vision impaired people .Research undertaken by Curtin University in conjunction with the Association for the Blind in Western Australia has shown that accessible e-learning environments can be developed to aid vision impaired adults achieve industry standard qualifications in IT networking. As computers become more ensconced in our private and business lives the need for useful IT knowledge rises. Industry standard training provides skills and knowledge for the vision impaired to maintain any computer network – their networks at the office as well as their home networks.

This paper describes an accessible e-learning environment designed to deliver advanced IT skills remotely, to legally blind students. The aim was to convert industry standard training written for the sighted into accessible formats for the vision impaired and deliver the learning materials in ways more suited to adult students with vision disabilities. The components of the learning environment, with particular emphasis on remote laboratory access, network topology graphics and simulators, are discussed together with the successes and problems faced in the hope that others may learn from our experience.

II. Scoping the Problem

Vision impaired (VI) adults continue to face problems in gaining employment. In the US the 2006 Disability Status Report (www.disabilitystatistics.org) reported an employment rate of only 47.5% for people with any sensory disability [2]. The 2002 Household Economic Studies reported a 55.3% employment rate for persons with communications disabilities, including vision impairment [3]. A further study on vision impaired youth employment levels reported a 28% employment rate for out-of-school youth [4,5] reports 25% of vision impaired in the UK are in employment and "younger people tend to be better qualified and there is a high correlation between qualification level and employment". Unemployment rate for vision impaired people in European countries in 2000 remained around 75% [6].

In each of these studies the employment figures for those with a vision disability were consistently lower than those for sighted individuals. Major contributors to this situation are suggested to be inability to access further education and the digital divide created by the emergence of computers [7,8]. This raises the question 'Can vision impaired adult learners gain equivalent grades to sighted learners if specialist education was accessible'? If so, such training would increase their employability, giving opportunities for financial independence and a more 'normal' lifestyle.

A study of factors relating to employment of vision impaired in Turkey reports that education, gender, age, marital status and Braille literacy were significant factors that predict the probability that a vision impaired individual would be employed [9]. The low rate of unemployment in the vision impaired population has been the focus of much research and the main barriers to employment are reported to be the lack of employment skills, transportation, housing and access to information [10,11,12,13]. Although education and training is not the sole answer to the problem, post-secondary education and training has also been found to be a significant factor for obtaining employment in numerous other studies across the globe [14,15].

Capella-McDonnall's [16] study of factors relating to the VI gaining competitive employment reports four significant factors; vocational education as a rehabilitation service resulting in an educational certificate or degree, having worked since the onset of the impairment, reason for applying for vocational rehabilitation related to obtaining a job and a high quality relationship between the counselor and the VI client. Capella-McDonnall [16], p312) states "the effect of completing an educational program is powerful because the odds of attaining competitive employment were more than nine times greater for those who obtained an educational certificate or degree compared to those who did not receive education as a service at all."

Access to higher education and training in specialized skills in preparation for employment is also restricted to those who suffer from vision disabilities. The situation faced by vision impaired students attempting to gain contemporary advanced IT training and education was investigated by the authors. Although attempts have been made to increase accessibility of their training materials the major providers such as Cisco, Microsoft, Oracle, etc. still fall short in providing a fully accessible on-line environment for the vision impaired. Our analyses resulted in the following list of problems:

- Lack of student mobility to attend classrooms and navigate around a large university campus.
- Location in remote areas where education and training services are not available
- Inability to see the whiteboard in classrooms, necessitating the lecturer to explain in narrative form the concepts being illustrated
- Sighted lecturers unaware of the needs of vision impaired students
- The inaccessibility of graphic and visual teaching and learning materials and limited access to textual materials
- Inability to access laboratory exercises and inability to carry out tasks without the assistance of a sighted person
- Inability to access written examination questions and answer in the traditional manner
- Inability to access interactive media, drag and drop, and similar electronic teaching tools
- Inability to access simulation software and common operating system such as Linux.

This posed a challenge, as several vision impaired students at Curtin University of Technology were studying Cisco technologies in their undergraduate degrees, and these problems were major hurdles to their completion of their courses of study. This project examined accessible alternatives to address each of these difficulties.

III. Remote laboratories in education

The past decade has shown an increase in the uptake of remote laboratories for the delivery of practical exercises and distance learning in electrical engineering and computer-based education. The discussion as to the viability and effectiveness of real laboratories, virtual laboratories and remote laboratories is active with no resolution reached thus far. Gustavsson [17] claims there is nothing that will replace synchronous learning through face-to-face interaction. This is a brave claim as it assumes a homogenous student group. Enthusiasts of hands-on learning propose that working with real equipment results in much more information and many more cues [18], however the difference between preference and effectiveness needs to be considered. The study by Corter and colleagues [18] comparing remote and hands-on laboratories reported more than 90% of the student respondents rated the remote labs to be comparable or better than the hands-on labs. Ma and Nickerson [19] argue that although automated (simulated and remote) laboratories allow professors to teach large student numbers, automation may remove the serendipity associated with traditional laboratory learning. The flexibility of remote laboratories enables students to utilise the laboratory in different locations and at numerous points in times [20], as well as those with special needs [21]. A comparison of remote, real and virtual laboratories by Nedic, Machotka and Nafalski [22] found that remote labs also offer students a tele-presence in the laboratory, the performance of experiments on real equipment, collaboration, learning by trial and error plus the opportunity to perform analysis on real experimental data.

The main characteristic of remote laboratories when compared to hands-on (also referred to as ‘real’), simulations and virtual laboratories, is that students obtain the data and learning experience by controlling environments separated by geographical distance. The majority of these environments are offered via the Internet for easy access by students. The perception of reality by the student is at the core [23,24] and the aim is to immerse the students into the learning experience so that no difference from physical presence is perceived. Colwell and Colleagues [21] argue that practical work is imperative for learners of science to develop both their conceptual and procedural understanding.

Vision impaired students use screen reading software to convert the text displayed on the screen into audio output, or screen magnification software. This use of assistive technologies necessitates a different design of educational materials for the vision impaired, and especially laboratory exercises to concrete the learning by experience.

However, the design and development of laboratory exercises in distance learning for vision impaired students using assistive technologies requires a careful consideration of accessibility issues. Accessible materials and environments require careful planning and deployment of navigation mechanisms, structure, content design and communication methods, and the approach needs to be highly learner- centered [25]. A key factor in the design of learning environments is that the environment should not be the cause of any unnecessary frustration to the student, and should include the facilities to permit easier interpretation of the material, and also support direct interaction enabling students to spend their full cognitive resources on the task rather than the interface [26].

The use of remote laboratories has contributed to the offering of advanced transgeographic education being an affective means of eradicating ethnocentrism, xenophobia and cultural divides [27]. The divides caused by disabilities could also be added to that list, provided the remote learning environments incorporate essential accessibility features. With the majority of on-line learning materials in the science and technologies incorporating a vast amount of vision driven features, barriers to learning are erected for vision impaired students. These are characterized by a predominance of graphics, images and animation in the presentation of learning materials. The presentation of e-learning materials ranges from highly textual with accompanying images through to computerized exercises and games incorporating a high concentration of visual features to assist comprehension. All these components provide accessibility problems for vision impaired students. A diagram or picture clearly illustrates the concepts being introduced to sighted students, however, rarely are there detailed explanations of the diagram or picture in the supporting text. Blind students cannot see these diagrams and students with acute vision disabilities also have great difficulty comprehending what is being taught. The inaccessibility of learning materials is highly evident in not only higher education environments, but also in on-line industry standard training courses.

IV. Cisco Academy for the Vision Impaired (CAVI)

Curtin University of Technology commenced offering the Cisco Network Academy Program to mainstream (sighted) students as part of the Bachelor of Technology (Computer Systems & Networking) degree program in 2002. Shortly after, four vision impaired students expressed an interest in entering the Bachelor of Technology program. These students faced significant problems with accessibility to the Cisco course on-line materials as much of these materials were not accessible to non-sighted users. Extending the Cisco courses to vision impaired students posed numerous teaching and learning challenges.

The Cisco Academy for the Vision Impaired (CAVI) has been delivering the Cisco Academy Programs to blind and vision impaired students since 2003, with up to 9 students per year from the local area. In 2007 the program was expanded to include students located in other parts of Australia and the U.S.A, with 25 vision impaired students enrolled in that year. In 2008 the enrolments of legally blind students exceeded 120, from countries including India, Sri Lanka, Canada, Egypt, Australia and the U.S.A. During the intervening period, a number of vision impaired students entered the Bachelor of Technology (Computer Systems & Networking) course at Curtin University with the total number of vision impaired students (in the CAVI program) increasing to 146 in the year 2008. In order to achieve the practical components of the CCNA courses, remote access to router and switch bundles was required. A description of the remote laboratory established for these students together with teaching tools developed and an accessible network simulation application follows.

A. The Mechanics of Delivering the Curricula

The Cisco curriculum is “media rich”, with much of the content delivered as Flash and interactive web pages. This style of delivery is often unsuitable for vision impaired persons. The arrangement of frames is inaccessible to screen review applications (speech output), but more importantly the curriculum relies heavily on visual keys to illustrate learning objectives. Several problems, not apparent to most sighted users, are also inherent in the curriculum design. The first problem is that the diagrams are extremely difficult to access or even explain to a person who has been blind since birth. The second problem is that the arrangement of frames and the lack of correct ALT labels (text equivalent buttons) add to the complexity of the presented material. The vision impaired students also advise they have no way of accessing the content of interactive sessions and find the supporting text confusing and misleading. To overcome these issues many supporting applications and documents were created but are beyond the scope of this document.

The CAVI program utilizes blind instructors to deliver the Cisco course materials with the support of a sighted teaching assistant. Blind instructors have first hand experience of the difficulties encountered by the vision impaired students and understand the most effective ways of presenting the materials. The classroom environment consists of a laboratory containing a network of PCs fitted with assistive technologies, routers and associated network equipment. Classes run two full days per week over the academic year. Local students physically attend classes and remote students log in (via the Internet) to a virtual classroom to listen to the lectures and participate in the tutorial exercises. The virtual classroom provides the facilities for students to talk to one another as well as communicate with the instructors, similar to a normal classroom environment. The lectures are recorded and made available as audio files on the project website along with other teaching materials for access by the students at any time.

It may be argued that the most difficult issue in delivering e-learning to blind students is that of explaining the meaning of graphical information. To overcome this issue, textual descriptions were created for all graphics used in the courses, including the curriculum, laboratory manuals and on-line exams. An example textual description is given in the excerpt below and refers to the diagram in Figure 1, graphic 2 of 4 in the text description.

Page has 4 graphics

Graphic 1 shows the segmentation with routers. There are four hubs and one router in the picture. Router is in the middle connected to hubs in four corners, three stations (PC) are connected to each hub.

Segmentation with routers provides:
More manageable, greater functionality, multiple activate paths
Smaller broadcast domain
Operates at layer 3

Graphic 2 show routers connected by WAN technologies. There are 10 routers in the graphic. Four of them are connected to each other in a square shape (each one corner of a square). If we name these routers from 1 to 4 clockwise starting from the left top, router number 1 is connected to router number 2 with ATM (Asynchronous Transfer Mode). Router number 2 is connected to router number 3 with T1/E1 and T3/E3. Router number 3 is connected to router number 4 with ATM. Router number 4 is connected to router number 1 with T1/E1 and T3/E3.Each of these four routers are connected to other routers.

Router number 1 is connected to two other networks one a cable modem via a router and other one X.25 via other router.

Router number 2 is connected to a dial-up modem via a router.

Router number 3 is connected to SDMS via a router and to xDSL via another router.

Router number 4 is connected to ISDN network via a router.

There are antenna signal to router number 3 and satellite signal to ISDN router connected to router number 4.

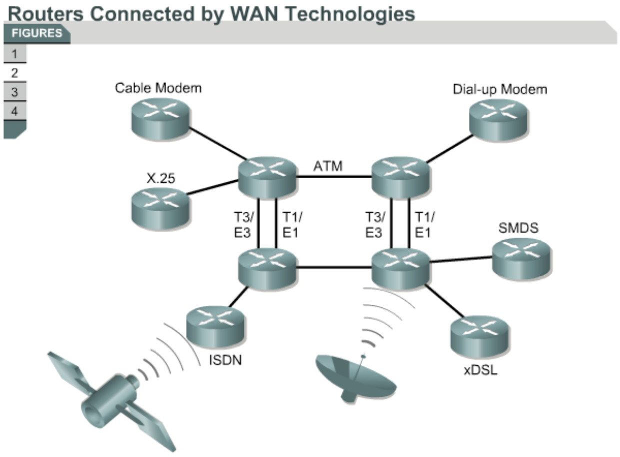


Figure 1: Example diagram from CCNA2 version 3.1 section 1.1.3

Further details of the teaching aids used in the project can be found in Murray [28] and on the project website http://www.cucat.org.

V. THE REMOTE LABORATORY USER EXPERIENCE

Laboratory exercises form a significant portion of the curriculum. In order for students situated remotely to access and participate in the laboratory sessions, a functional, remotely accessible network topology was developed. The configuration illustrated was developed for the CCNA version 3.1 curriculum, with work currently underway to reconfigure for Discovery and Exploration curricula to be delivered in 2009. Laboratory equipment generally consists of three routers and two switches. The configuration may be described as two branch offices, say Perth and Sydney, connected together via the ISP or Internet cloud. The edge routers are configured by the students to allow connectivity via the middle router (cloud or ISP). Local students interact directly with the routers' configuration via serial (console) interfaces. A problem exists when attempting to allow remote students access to "real" routing hardware. The routers may not be placed on production networks for obvious reasons and initial configurations must be entered via the console connection. Therefore requirements for a remote lab must allow students to perform:

- Initial configuration via the console cable
- Remote power cycling of network equipment and workstations
- Connectivity tests
- Advanced router and switch configuration.

One such system does exist, Netlab, developed and distributed by NDG (http://www.netdevgroup.com/), however the cost of this system is a major factor hindering its adoption. Additionally, the java based applications in Netlab, including the booking system, telnet client to interact with the network hardware and server system are not accessible by screen readers (software utilized by blind computer users to convert on-screen information to audio or Braille output). The CAVI system developed costs significantly less than the Academy Edition of Netlab. Whilst it does not offer advanced features such as equipment booking it performs all the required functions for the vision impaired class applications. In its most simple form, it consists of the standard CCNA laboratory bundle: 3 routers and 2 switches, with several virtualized Linux PC servers running FTP, HTTP, Telnet and other associated services; all may be accessed by their serial ports (see Figure 2).

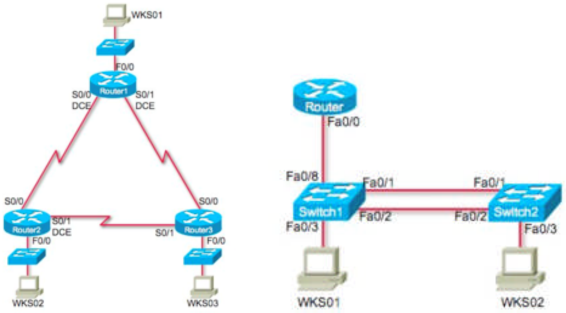


Figure 2: Standard router bundle (top) & Switch Bundle (bottom)

A standard serial port or console switch is used to access the devices in the laboratory bundle. Remote students may telnet into the console server (a device that allows Ethernet to multiple serial port connections), accessing the routers, switches and Linux servers from any locality worldwide. The use of Linux on the host and server machines is necessary as the command line may be accessed through the serial ports allowing the students to connect directly to the server hosting the multiple virtual machines. Users may then telnet to the virtual machines and access the command line via the screen reader. Virtualized GUI based operating systems are not easily accessible to the assistive technology when installed behind the console switch.

The physical layout of the remote laboratory equipment is depicted in Figure 3.

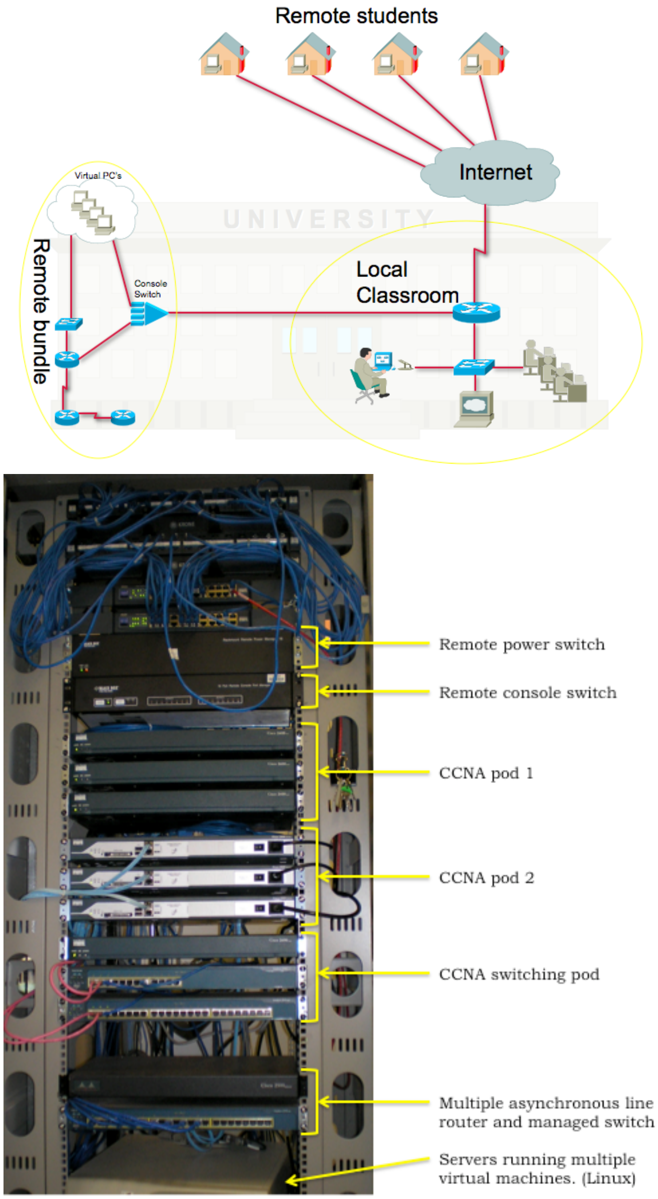
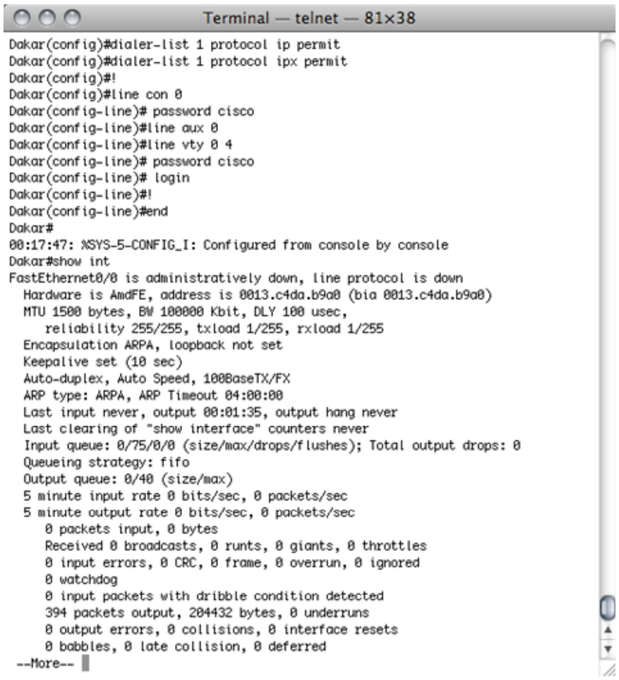


Figure 3: Teaching environment (top) and physical remote laboratory equipment layout (bottom)

Figure 4 illustrates an active telnet session logged into the remote bundle. As the routers are on their own network, with remote access attaching only to the serial ports, this system does not offer any security risk to the institution utilizing it. Once the student has authenticated with the console switch (simple plain text password) a list of available equipment is displayed, as shown in Figure 5.



```
Dakar(config)#dialer-list 1 protocol ip permit
Dakar(config)#dialer-list 1 protocol ipx permit
Dakar(config)#!
Dakar(config)#line con 0
Dakar(config-line)# password cisco
Dakar(config-line)#line aux 0
Dakar(config-line)#line vty 0 4
Dakar(config-line)# password cisco
Dakar(config-line)# login
Dakar(config-line)#!
Dakar(config-line)#end
Dakar#
00:17:47: %SYS-5-CONFIG_I: Configured from console by console
Dakar#show int
FastEthernet0/0 is administratively down, line protocol is down
  Hardware is AmdFE, address is 0013.c4da.b9a0 (bia 0013.c4da.b9a0)
  MTU 1500 bytes, BW 100000 Kbit, DLY 100 usec,
     reliability 255/255, txload 1/255, rxload 1/255
  Encapsulation ARPA, loopback not set
  Keepalive set (10 sec)
  Auto-duplex, Auto Speed, 100BaseTX/FX
  ARP type: ARPA, ARP Timeout 04:00:00
  Last input never, output 00:01:35, output hang never
  Last clearing of "show interface" counters never
  Input queue: 0/75/0/0 (size/max/drops/flushes); Total output drops: 0
  Queueing strategy: fifo
  Output queue: 0/40 (size/max)
  5 minute input rate 0 bits/sec, 0 packets/sec
  5 minute output rate 0 bits/sec, 0 packets/sec
     0 packets input, 0 bytes
     Received 0 broadcasts, 0 runts, 0 giants, 0 throttles
     0 input errors, 0 CRC, 0 frame, 0 overrun, 0 ignored
     0 watchdog
     0 input packets with dribble condition detected
     394 packets output, 204432 bytes, 0 underruns
     0 output errors, 0 collisions, 0 interface resets
     0 babbles, 0 late collision, 0 deferred
 --More--
```

Figure 4: Active console session

```
[Murray-2:~] iainmurray% telnet 134.7.43.171
Trying 134.7.43.171...
Connected to 134.7.43.171.
Escape character is '^]'.

Enter Password: ********

PORT STATUS:  Version 3.0,  Site ID: Curtin -o o- --oo o- -o oo -o

PORT |     NAME      |  PASSWORD    | STATUS | MODE | BUFFER
COUNT
-----+------------------+------------------+--------+--------+--------------
 09 | Router1     | (defined)    | Free | Any  |     0
 10 | Router2     | (defined)    | Free | Any  |     0
 11 | Router3     | (defined)    | Free | Any  |     0
 12 | WKS1        | (defined)    | Free | Any  |     0
 13 | WKS2        | (defined)    | Free | Any  |     0
 14 | WKS3        | (defined)    | Free | Any  |     0
```

Figure 5: Remote bundle equipment list

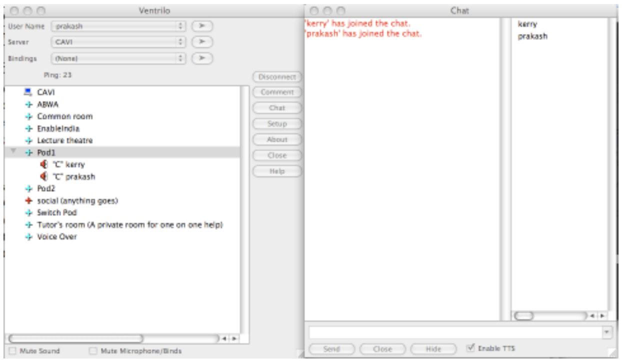
Several commands are available and are listed in Table 1. Connection to equipment is made via the /C n command, where n = required equipment port number.

TABLE I. CONSOLE SWITCH COMMAND MENU

Display Options	
/S /SD	Port Status
/W	Port Parameters (who)
/J	Site ID
/H	Command Menu (Help)
Control	
<Enter>	Enter Command Mode
/x	Exit Command Mode
/C n	Connect to Port (n: Port# or name)

As the booking system was incomplete at the time of writing, a virtual classroom was utilized as a method of ensuring students knew if the equipment was in use. When undertaking a laboratory, students logged into the Ventrilo server (a voice communication application designed for on-line gamers) and entered the appropriate channel, as shown in Figure 6. In this way students not only can tell if a particular bundle is in use but may also conduct laboratory sessions collaboratively with other students.

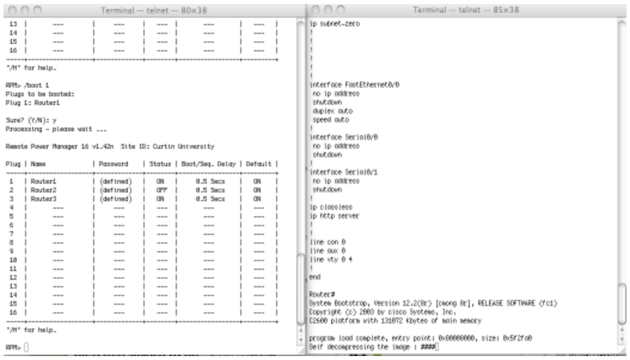
Power cycling of equipment is undertaken by authenticating first to the console switch and connecting to the remote power switch. The power switch may then be used to power down individual devices within the bundle. This is usually done with a secondary telnet session, allowing access to the router/switch to be power cycled and therefore the boot process to be interrupted (as in the case of password recovery laboratories). Each device may be powered on, off or rebooted. Figure 7 illustrates the process of remotely rebooting a router.



User Name: prakash
Server: CAVI
Bindings: (None)
Ping: 23

'kerry' has joined the chat.
'prakash' has joined the chat.

Figure 6: Ventrilo session (note users in the “Pod 1” channel)



```
 13  |     ---      |    ---     |  --- |      ---       |  ---  |
 14  |     ---      |    ---     |  --- |      ---       |  ---  |
 15  |     ---      |    ---     |  --- |      ---       |  ---  |
 16  |     ---      |    ---     |  --- |      ---       |  ---  |
-----+--------------+------------+------+----------------+-------+
"/H" for help.

RPM> /boot 1
Plugs to be booted:
Plug 1: Router1

Sure? (Y/N): y
Processing - please wait ...

Remote Power Manager 16 v1.42n  Site ID: Curtin University

Plug | Name         | Password   |Status| Boot/Seq. Delay|Default|
-----+--------------+------------+------+----------------+-------+
 1   | Router1      | (defined)  |  ON  |    0.5 Secs    |  ON   |
 2   | Router2      | (defined)  |  OFF |    0.5 Secs    |  ON   |
 3   | Router3      | (defined)  |  ON  |    0.5 Secs    |  ON   |
 4   |     ---      |    ---     |  --- |      ---       |  ---  |
 5   |     ---      |    ---     |  --- |      ---       |  ---  |
 6   |     ---      |    ---     |  --- |      ---       |  ---  |
 7   |     ---      |    ---     |  --- |      ---       |  ---  |
 8   |     ---      |    ---     |  --- |      ---       |  ---  |
 9   |     ---      |    ---     |  --- |      ---       |  ---  |
 10  |     ---      |    ---     |  --- |      ---       |  ---  |
 11  |     ---      |    ---     |  --- |      ---       |  ---  |
 12  |     ---      |    ---     |  --- |      ---       |  ---  |
 13  |     ---      |    ---     |  --- |      ---       |  ---  |
 14  |     ---      |    ---     |  --- |      ---       |  ---  |
 15  |     ---      |    ---     |  --- |      ---       |  ---  |
 16  |     ---      |    ---     |  --- |      ---       |  ---  |
-----+--------------+------------+------+----------------+-------+
"/H" for help.

RPM>
```

```
ip subnet-zero
!
!
!
!
!
!
interface FastEthernet0/0
 no ip address
 shutdown
 duplex auto
 speed auto
!
interface Serial0/0
 no ip address
 shutdown
!
interface Serial0/1
 no ip address
 shutdown
!
ip classless
ip http server
!
!
line con 0
line aux 0
line vty 0 4
!
end

Router#
System Bootstrap, Version 12.2(8r) [cmong 8r], RELEASE SOFTWARE (fc1)
Copyright (c) 2003 by cisco Systems, Inc.
C2600 platform with 131072 Kbytes of main memory

program load complete, entry point: 0x80008000, size: 0x5f2fa0
Self decompressing the image : ####
```

Figure 7: Remotely rebooting the router

Once connected via telnet, students may configure routers, switches and workstations in the same manner as if physically present. One such session is shown in Figure 7. Although this system allows laboratories to be completed, including e-Labs it has several shortcomings. These include the lack of a formal booking system, leading to students attempting to access the limited resources while others are engaged in laboratory sessions and it does not give students experience with the physical cabling of network systems. To assist students with understanding the physical aspects of cabling, recorded audio demonstrations of the physical features of cables, connectors and their locations on switch and router hardware are conducted. In these demonstrations, a local vision impaired student conducts a supervised cable lab, describing in detail what they "feel" and how connections are made, in much the same manner as video is utilized for sighted students.

VI. iNetSim Network Simulator for Apple OS X

iNetSim is a accessible network simulator, created to allow both vision-impaired and sighted users to complete CCNA 2 laboratory sessions without access to the networking hardware [30]. Existing software used in the CCNA course for network simulation and laboratory practice (Packet Tracer) and the eLabs is not accessible to those with impaired vision as it utilizes images of network topology, allows only mouse selection of network devices and tools and is incompatible with screen reading software. In contrast, iNetSim has been developed to be accessible by blind and vision impaired users in addition to those with normal vision. All user interface and network topology elements are accessible via the Apple screen reader (VoiceOver) keyboard shortcuts and provide a meaningful response when read by VoiceOver. Network simulators usually rely on the use of a mouse to add simulated communication links between devices, place network devices on the work area canvas, select configuration options and view simulation results. To connect two devices with a communications link, the user must generally click on icons for the simulated devices and drag the connection to its end point, another network device under normal circumstances. As this is usually not possible for vision-impaired users, iNetSim also incorporates the use of tables for connecting devices. Tables are used to alter a device's location in the topology area, and configure ports and links. Tables are used as navigation with speech prompts, as these can be accessed with VoiceOver shortcut keys and cursor keys. iNetSim can be used solely with the keyboard, therefore the eye and hand issues faced by vision-impaired students can be avoided. As a GUI is also available, sighted iNetSim users can alternatively use a more traditional drag-drop mouse-based interface.

The system is capable of representing several generic network devices including routers, switches, hubs and PCs. Each device must be configured via a command line for correct operation. Figure 8 illustrates the application running with the textual command line terminal session to Router0 open. Note that the IP address on Router0, interface S1 is set to the same value as in the port table (highlighted) under the main canvas. Selecting values from the drop down boxes or edit fields in the main application window has the same effect as entering the command line configuration. Changes made with either method will be reflected throughout the application. This allows rapid basic configuration to be undertaken by the instructor so that students may concentrate on the particular task in the session.

Each device may have several ports of different types including Ethernet, serial and console. The user creates a connection by specifying two ports to connect and a cable type. Removing a connected port disables the connection the same way unplugging a cable would in a real network. The command line interface to devices also provides control and feedback over the simulation. The interface acts in a similar way to the operating system for that device type (e.g. a generic DOS-like system for PCs and Cisco IOS for routers). A subset of the commands applicable to CCNA 2 allows the user to display and modify device configuration, establish routing protocols and ping, Traceroute or telnet to other devices. iNetSim maintains a representation of routing tables to simulate these tasks correctly.

A completed laboratory is depicted in the screen capture illustrated in Figure 8. Note the configuration entered in the terminal screen matches the configuration in the tables and the successful pings from both the routers and workstations.

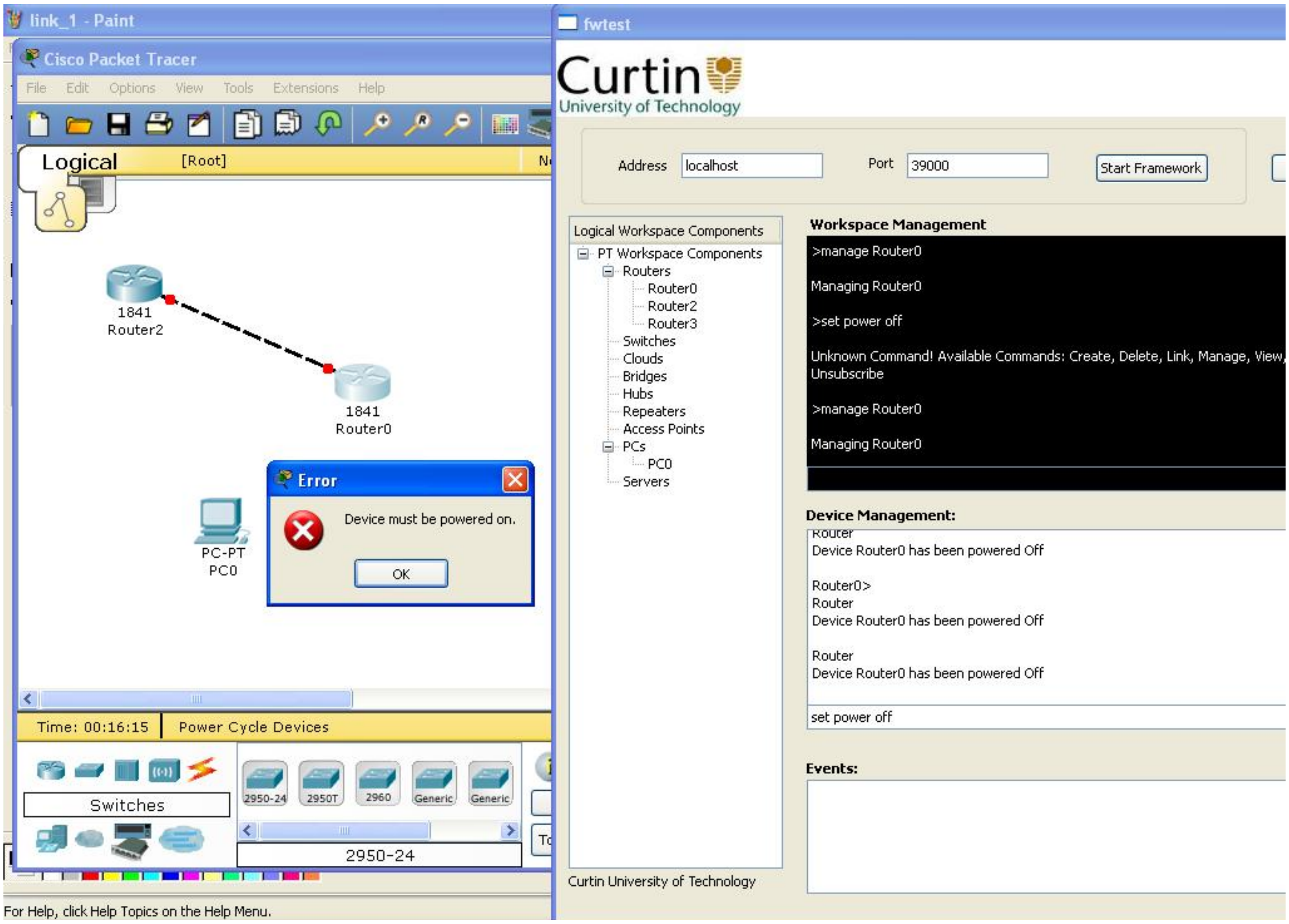
A. *Packet Tracer Accessibility*

iNetSim was successful in its aim of illustrating that network simulators may be made fully accessible if accessibility is built into the design stage of application development. However, maintaining and building a separate simulator for use by the vision impaired is not feasible. Any such application would lag development and features of commercial, well resourced projects such as Packet Tracer, hence it was decided to develop an external application to connect to Packet Tracer utilizing the newly released APIs and multi user features. Initial development has aimed at examining the flexibility and ability of the framework underlying the Packet Tracer application development environment. The information obtained was then used to decide on the feasibility of an external application for vision impaired people, allowing them to use and manipulate the Packet Tracer software package to carry on networking simulation, particularly where the CCNA curriculum is concerned.

Although Packet Tracer is not intended as a substitute for real equipment, it allows students to practice using a model of the Cisco Internetwork Operating System (IOS) command line interface and provides visual, drag-and-drop problem solving using virtual networking devices. This hands-on capability is a fundamental component of learning how to configure routers and switches from the command line. Students can see how to configure and connect networking hardware while confirming systems design. Instructors can create their own self-evaluated activities that present immediate feedback to students on their proficiency in completing assignments.

Starting from version 5.0 onwards, Packet Tracer supports external applications as well as user connections. In the case of multiuser connections, an instance of Packet Tracer on computer A can communicate with another instance of Packet Tracer on computer B. Therefore, users in different places can have collaborative and competitive network building, using the real network to carry the virtual packets.

Based on Packet Tracer Messaging Protocol (PTMP) and Inter Process Communication (IPC) the external application developed communicates with an established Packet Tracer instance on the same computer or another on the network, thus it is only necessary that the developers maintain an accessible client that communicates with the commercial simulation software.



Device must be powered on.

Address localhost

Port 39000

```
>manage Router0

Managing Router0

>set power off

Unknown Command! Available Commands: Create, Delete, Link, Manage, View,
Unsubscribe

>manage Router0

Managing Router0
```

```
Router
Device Router0 has been powered Off

Router0>
Router
Device Router0 has been powered Off

Router
Device Router0 has been powered Off
```

set power off

Figure 8: External application communicating with an instance of Packet Tracer

VII. Packet Tracer Exercises and eLabs

The combination of access to real network hardware and a practice environment in the simulation software has shown to be a valuable resource. CAVI has converted Discovery Packet Tracer exercises and eLabs (from CCNA version 3.1) to a format that may be run on the remote bundles so that students receive a similar learning experience to that of their sighted peers. These conversions include descriptions of graphics and network topologies, starting router configurations that are pasted into the routers and accessible instructions on laboratory procedure. eLabs are utilized to communicate a single learning outcome from the curriculum chapter being currently studied without the necessity of the student creating multiple complex router configurations. Vision impaired students took significantly longer than their sighted counterparts to complete each eLab due to the necessity of connecting to the remote bundle, pasting configurations into the required network prior to commencing the laboratory session. However, students found that completing a set of eLabs reduced this setup time considerably. Trials of eLabs with instructors, both vision impaired (n=3) and sighted (n=4) established that there was no perceived difference in the user experience once configurations were copied to laboratory equipment. On a scale of 0-5 where 0 = of no use, 5 = very useful, students (n=13) rated the usefulness of eLabs at an average of 4.6 (median = 4.5). The disadvantage of eLabs is primarily in the preparation involved, each piece of equipment used in a particular session required a configuration file to be created and tested. Instructions needed to be transcribed from the Flash files, tested by qualified personnel and altered to suit use on the remote bundle. In excess of 180 individual files were created. A further disadvantage is in cases where a large number of network devices are required, simplified topologies were used due to restrictions on the quantity of network devices in the remote bundle.

VIII. CONVEYING GRAPHICAL NETWORK TOPOLOGIES TO THE BLIND

A. *Network Dominoes*

The concept of a tactile method of displaying network devices and interconnections was initially well received by students in the pilot study. A survey of totally blind users of the Network Dominoes (see Figure 9) showed that students rated the usefulness highly at 4.4 out of 5 (n=6). An interesting item was raised in the comments section of the survey, as detailed in the following quote.

> *"The network dominoes are interesting for showing students what the network shapes look like. This can be useful for if they have a sighted person without any networking knowledge trying to explain a diagram to them, they are able to tell the sighted person what the shapes are."*

The ability to identify standard graphic icons is of obvious import given that it is expected that students will be in mainstream employment and required to interact with their sighted peers. Whilst the network dominoes achieved the desired result of communicating topologies, it was decided to discontinue their use in the later iterations of the courses. The primary concern with the network dominoes became apparent with the inclusion of students undertaking the courses remotely. Remote students would connect the dominoes in the manner that they thought was correct, but on occasion was significantly different to the intended topology. As the instructors have no method of checking the said topology, being both blind and geographically separated from the tactile topology, errors in construction would not be identified and thus lead to possible misconceptions by students. A secondary issue was that of cost, both in production of the full range of network device objects and that of distribution of multiple sets. If the cost of production could be reduced, use of these devices should be explored in future trials.

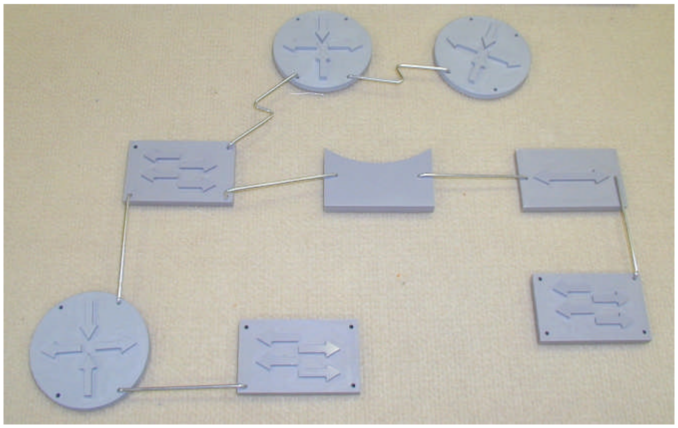


Figure 9: Network Dominoes

B. *Tactile Diagrams*

PIAF, Picture in a Flash, is an assistive technology device that uses a controlled heat source to imprint images or diagrams onto heat sensitive paper which then swells to reproduce a raised representation of the image. The output of PIAF is intended to allow the vision impaired to see by feeling tactile graphics. While in many situations this product provides a satisfactory solution to gaining an appreciation of an image or basic diagram it does not meet the requirements necessary to impart understanding of detailed information taken from complex technical diagrams. As Dulin [29] identified, raised line drawings may increase the blind individual's spatial cognition and communicate information from graphics that would otherwise be inaccessible, several issues were noted when utilizing this media in the context of technical drawings. Many of the network diagrams, in order to fit on the A4 capsule paper, required the network device icons to be of limited dimensions (approximately 30mm by 30mm) the tactile resolution of the human senses made it difficult to identify or differentiate between similar objects, for example hubs and switches. A significant number of tactile diagrams were produced, utilizing the PIAF system, in the pilot study stage of the research (in excess of 150 individual diagrams and charts). Totally blind students within the pilot study were surveyed (n=6) on the suitability and usefulness of this style of graphic representation with the disappointing results, given the cost and time taken in production. Overall on the scale of 0 to 5 an average result of 2.4 (mean of 2.0) was returned. The physical bulk and material cost of tactile pictures made it difficult and inefficient to distribute up-to-date material to remote students, particularly when compared to electronic text descriptions.

PIAF along with the tactile assistive technology devices suggested as being useful to vision impaired students proved to be unsatisfactory in meeting the needs of the vision impaired student studying to the Cisco course materials. The complexity of images and diagrams in the required teaching materials and the individual health issues of students proved these devices/methods unsatisfactory. Given the logistical difficulties of shipping significant quantities of tactile pictures and overlays to various countries the costs become prohibitive as the class size increases.

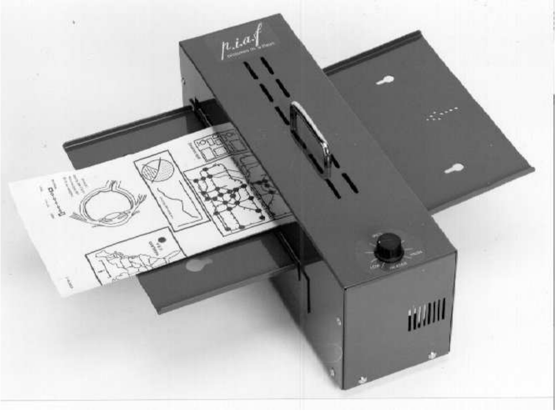
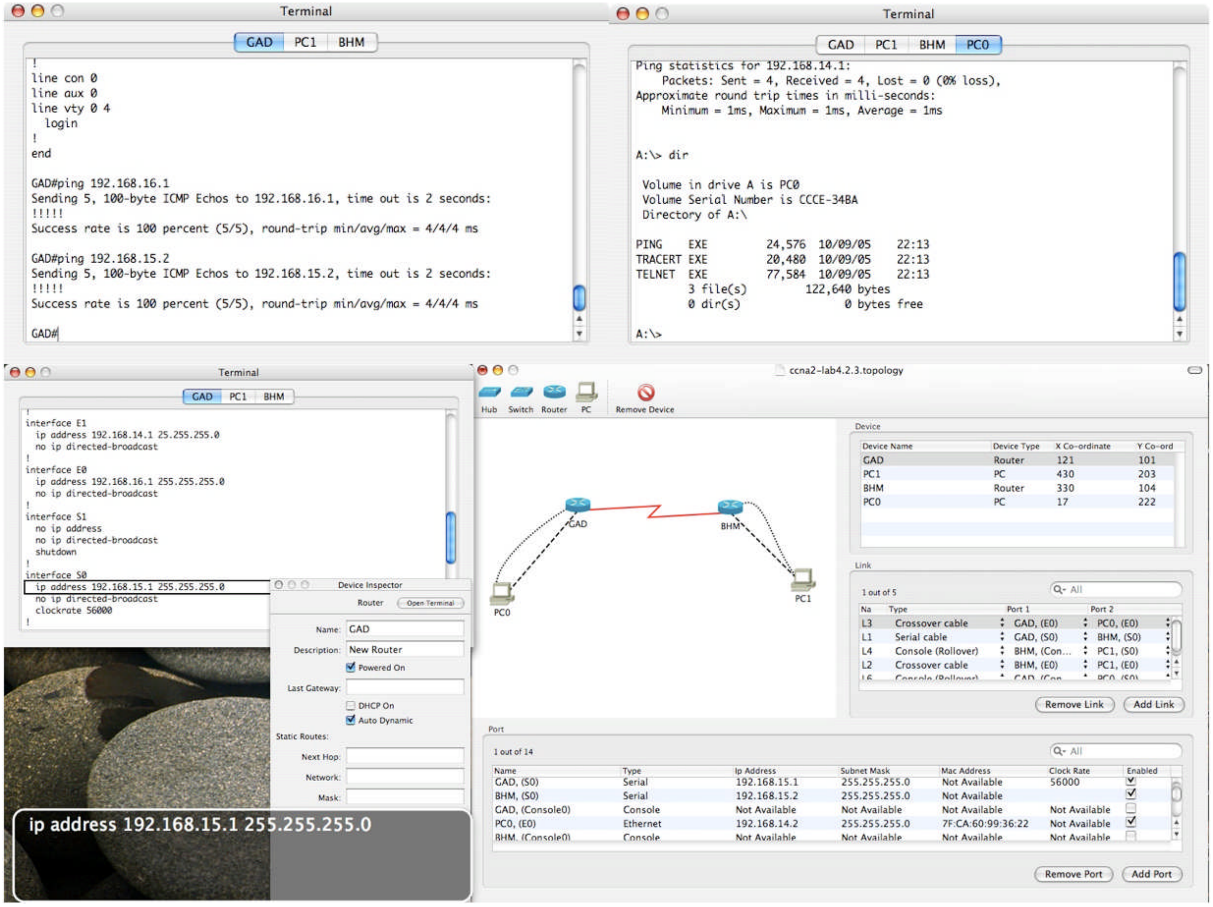


Figure 10: PIAF Tactile Printer and Nomad (Source: http://www.brailleworldindia.com/braille__tactile_graphics.htm#PIAF)

The challenges relating to teaching materials focused directly on how best the students could gain an understanding of course materials. While not all, many complex computer related technical images and diagrams rely on the use of color, which is difficult to represent on tactile media. The main difficulty with these devices or tactile media was that the demands being placed on them were inconsistent with their ability to meet teaching requirements.



```
!
line con 0
line aux 0
line vty 0 4
  login
!
end

GAD#ping 192.168.16.1
Sending 5, 100-byte ICMP Echos to 192.168.16.1, time out is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 4/4/4 ms

GAD#ping 192.168.15.2
Sending 5, 100-byte ICMP Echos to 192.168.15.2, time out is 2 seconds:
!!!!!
Success rate is 100 percent (5/5), round-trip min/avg/max = 4/4/4 ms

GAD#
```

```
Ping statistics for 192.168.14.1:
    Packets: Sent = 4, Received = 4, Lost = 0 (0% loss),
Approximate round trip times in milli-seconds:
    Minimum = 1ms, Maximum = 1ms, Average = 1ms


A:\> dir

 Volume in drive A is PC0
 Volume Serial Number is CCCE-34BA
 Directory of A:\

PING     EXE        24,576  10/09/05    22:13
TRACERT  EXE        20,480  10/09/05    22:13
TELNET   EXE        77,584  10/09/05    22:13
         3 file(s)        122,640 bytes
         0 dir(s)               0 bytes free

A:\>
```

```
!
interface E1
  ip address 192.168.14.1 25.255.255.0
  no ip directed-broadcast
!
interface E0
  ip address 192.168.16.1 255.255.255.0
  no ip directed-broadcast
!
interface S1
  no ip address
  no ip directed-broadcast
  shutdown
!
interface S0
  ip address 192.168.15.1 255.255.255.0
  no ip directed-broadcast
  clockrate 56000
!
```

Name: GAD
Description: New Router
☑ Powered On
Last Gateway:
☐ DHCP On
☑ Auto Dynamic
Static Routes:
Next Hop:
Network:
Mask:

ip address 192.168.15.1 255.255.255.0

Device

Device Name	Device Type	X Co-ordinate	Y Co-ord
GAD	Router	121	101
PC1	PC	430	203
BHM	Router	330	104
PC0	PC	17	222

Link

1 out of 5

Na	Type	Port 1	Port 2
L3	Crossover cable	GAD, (E0)	PC0, (E0)
L1	Serial cable	GAD, (S0)	BHM, (S0)
L4	Console (Rollover)	BHM, (Con...	PC1, (S0)
L2	Crossover cable	BHM, (E0)	PC1, (E0)
L6	Console (Rollover)	GAD, (Con...	PC0, (S0)

Port

1 out of 14

Name	Type	Ip Address	Subnet Mask	Mac Address	Clock Rate	Enabled
GAD, (S0)	Serial	192.168.15.1	255.255.255.0	Not Available	56000	☑
BHM, (S0)	Serial	192.168.15.2	255.255.255.0	Not Available		☑
GAD, (Console0)	Console	Not Available	Not Available	Not Available	Not Available	☐
PC0, (E0)	Ethernet	192.168.14.2	255.255.255.0	7F:CA:60:99:36:22	Not Available	☑
BHM, (Console0)	Console	Not Available	Not Available	Not Available	Not Available	☐

Figure 11: iNetSim with command line terminal session open. The text in the floating window shows the text read by VoiceOver

IX. Student Results

The following discussion covers an analysis of how effectively and efficiently problems defined in relation to remote teaching of vision impaired students were addressed. [29] Definitions of effectiveness and efficiency as the basis for these evaluations are:

- Effectiveness is evaluated by determining how well the solution achieves the given objectives. In this context effectiveness is measured by determining whether the vision impaired students studying the converted e-learning courses achieved the same outcomes as able-bodied students studying the unconverted courses.
- Efficiency is evaluated by determining any increase or decrease in the level of the resources used to achieve the stated objectives. In this context efficiency is measured by ascertaining whether the difference in costs for teaching methods and tools between the vision impaired accessible environment and the traditional Cisco e-learning environment is minimal for the same level of output.

Each mode of presentation is evaluated via a ranking of 0 through 5, where 0 denotes “no use whatsoever”, 3 gives the same or similar outcomes or resources, and 5 shows a significant increase in outcomes or significant decrease in resources. An acceptable solution should rate at least 3 and a summary of the ratings is presented in Table 2.

TABLE II. EFFECTIVENESS OF SOLUTIONS

Requirements	Potential Solution	Effectiveness (same outcomes)	Efficiency (more or less resources)	Comments
1. Lack of student mobility	Virtual classroom	4	5	Highly scalable without increasing resources
	Remote Bundle	5	4	Slightly more complex to use. Reduces cost of distributing multiple laboratory pods. Equipment available 24/7
2. Inclusion of remote students	Virtual classroom	4	5	No realistic limitations on lecture size
7. Cannot access laboratory exercises	e-Labs	3	3	Labour intensive conversion process.
	Remote labs	5	4	No access to physical cabling. 24/7 availability of equipment. Lower cost to small Academies (shared resource)
	iNetSim	3	2	Proof of concept only.
	Network dominos	3	2	High production cost and bulk.
10. network topology/ simulation software	Network dominos	2	2	High production cost and bulk. Not interactive and do not simulate the operating system.
	iNetSim	3	2	Modification of Packet Tracer user interface to overcome the access issues.

Overall results for students in the 2007 intake are shown in Table 3 and **Figure 12**.

TABLE III. STUDENT RESULTS

Course	Demographic	n	Comments
CCNA 1	Population	214,499	Average across all questions
	Vision Impaired	24	
	Sighted	19	
CCNA 2	Population	67,601	Average across all questions
	Vision Impaired	23	
	Sighted	19	
ITE 1	Population	61,386	Average across all questions
	Vision Impaired	23	
	Sighted	19	
ITE 2	Population	28,484	
	Vision Impaired	22	
	Sighted	32	

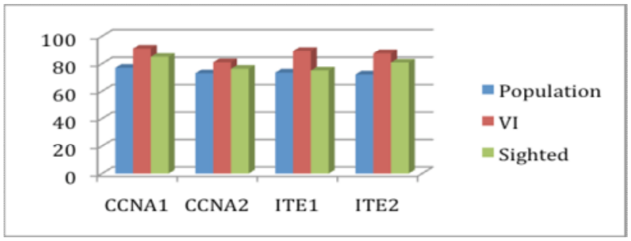


Figure 12: Student Examination Results

As can be noted by the graph in **Figure 12**, the vision impaired group consistently scored better across the full range of courses when compared to the sighted groups. The population group may not be representative as no level of support, availability of resources and other relevant information is available for this group. The sighted group may be directly compared as they consisted of students taught at Curtin University of Technology as part of their coursework in the Bachelor of Technology (Computer Systems and Networking) and as such had a comparable educational background and access to resources such as laboratory equipment and information.

X. CONCLUSIONS

Whilst the use of the remote bundle overcomes many of the limitations imposed by delivering CCNA laboratories to remote blind and vision impaired students, further work is required to improve functionality and ease of use. A method of booking and authentication of users is currently under development as part of a "front end" web portal to streamline the connection to network devices.

iNetsim was successful in proving that network simulators may be made accessible. However, it is now considered that, with the availability of Packet Tracer API's and the Packet Tracer Messaging Protocol a possible way forward in accessibility is to develop an accessible extension (a user interface) that communicates with Packet Tracer. This would have the benefit of utilizing the superior Packet Tracer library of devices and protocols and the support and continuing development without duplication of resources.

The solutions presented in this paper assist in overcoming the laboratory issues involved in remote delivery to vision impaired students however there are many significant obstacles in accessibility that have also been addressed in the CAVI project [31]. The CAVI classes offer a holistic environment tailored to cater for blind and low vision students without compromising course quality and student outcomes. Using the environment and tools established by CAVI students with severe vision impairment are able to undertake the same Cisco courses as their sighted counterparts.

REFERENCES

[1] Murray, I. and Armstrong, H. 2009. Remote Laboratory Access for Students with Vision Impairment. In Proceedings of the 2009 Fifth international Conference on Networking and Services - Volume 00 (April 20 - 25, 2009). ICNS. IEEE Computer Society, Washington, DC, 566-571. DOI= http://dx.doi.org/10.1109/ICNS.2009.107

[2] Rehabilitation Research and Training Centre on Disability Demographics and Statistics (2007). 2006 Disability Status Report. Ithaca, NY. Cornell University. Available at: http://www.disabilitystatistics.org

[2] Steinmetz, E. (2006) Americans with Disabilities: 2002 Household Economic Studies, Current Population Reports (issued May 2006), Dept Commerce, U.S. Census Bureau. Available at: http://www.census.gov/prod/2006pubs/p70-107.pdf

[4] Wagner, M., Newman, L., Camelo, R., Garza, N & Levine, P. (2005) after high school: a first look at the postschool experiences of youth with disabilities. A report from the National Longitudinal Transition Study -2 (NLTS2) Menlo Park, CA, SRI International. Available at www.nlts.org/reports/2005_04/nlts2_report_2005_04_complete.pdf

[5] Dryden, G. (2000). Training, rehabilitation and employment for visually impaired people in the UK, Royal National Institute of the Blind, Available at: http://www.euroblind.org/fichiersGB/emploidryden.htm

[6] Osoian, C. Zaharie, M & Stegerean, R. 2008, Overcoming Barriers to employment for visual impaired persons in the Romanian labor market, Studia Universitatis Babes Bolyai – Oeconomica, Issue 1, pages 34-44

[7] Department of Training & Employment. (2000). Building Diversity Project 2000. Western Australian Department of Training and Employment, Government of Western Australia.

[8] Hollier, S., (2007). The Disability Divide: A Study into the Impact of Computing and Internet-related Technologies on People who are Blind or Vision Impaired, PhD Thesis, Curtin University of Technology, Perth, Western Australia

[9] Bengisu, M., Izbirak, G. & Mackieh, A., 2008. Work-Related Challenges for Individual who are Visually Impaired in Turkey, Journal of Visual Impairment & Blindness, 102, 284-294

[10] Crudden, A. & McBroom, L.W., 1999. Barriers to employment: A survey of employed persons who are visually impaired, Journal of Visual Impairment & Blindness, 93, 341-350

[11] Crudden, A., Sansing, W. & Butler, S., 2005. Overcoming barriers to employment: Strategies of rehabilitation providers, Journal of Visual Impairment & Blindness, 99, 325-334

[12] Moore, J.E. & Wolffe, K.E., (1997), Employment considerations for adults with low vision, In A.L. Corn & A.J. Koenig (Eds)., Foundations of Low Vision: clinical and Functional Perspectives, pp 340-362, AFB Press, New York

[13] O'Day, B., (1999). Employment Barriers for people with visual impairments, Journal of Visual Impairment & Blindness, 93, 627-642

[14] Kirchner, C., Schmeidler, E., & Todorov, A. (1999). Looking at employment through a life span telescope: Age, health and employment status of people with serious visual impairment, Mississippi State: Mississippi State University Rehabilitation Research and Training Center on blindness and Low Vision.

[15] Lee, I.S., & Park, S.K., (2008). Employment Status and Predictors Among People with Visual Impairments in South Korea: Results of a National Survey, Journal of Visual Impairment & blindness, 102, 147-159

[16] Capella-McDonnall, M.E., 2005. Predictors of competitive Employment for Blind and Visually Impaired consumers of Vocational Rehabilitation Services, Journal of Visual Impairment & Blindness, 99, 303-315

[17] Gustavsson, I, Remote Laboratory Experiments in Electrical Engineering Education, Fourth IEEE International Caracas Conference on Devices, Circuits and Systems, Aruba, IEEE Press, April 2002

[18] Corter, J.E. Nickerson, .J.V, Esche, S.K. and Chassapis, C, Remote Versus Hands-On Labs: A Comparative Study, 34th ASEE/IEEE Frontiers in Education Conference, Savannah, GA, IEEE Press, October 2004

[19] Ma, J. and Nickerson, J.V. Hands-On, Simulated and remote Laboratories: A comparative Literature Review, ACM Computing Surveys, vol. 38, no. 3, article 7, September 2006

[20] Canfora, G, Daponte, P. and Rapuano, S. Remotely Accessible Laboratory for Electronic Measurement Teaching, Computer Standards and Interfaces, vol. 26, no. 6, 2004, pp 489-499

[21] Colwell, C, Scanlon, E and Cooper, M Using Remote Laboratories to Extend Access to Science and Engineering, Computers & Education, vol. 38, 2002, pp 65-76

[22] Nedic, Z, Machotka, J and Nafalski, A, Remote laboratories versus Virtual and Real laboratories, 33rd ASEE/IEEE Frontiers in Education, Boulder, CO, 2003

[23] Bentley, F., Tollmar, O, Demirdjian, D. , Oile, K. and Darrell, T. Perceptive Presence, IEEE Computer Graphics and Applications, vol 23, no. 5, 2003, pp 26-36

[24] Biocca, F., Inserting the Presence of Mind into a Philosophy of Presence: A response to Sheridan and Mantovaniand Riva, Presence: Teleoperators and Virtual Environments, vol 10, no. 5, 2001, pp 546-556

[25] Pearson, E.J. and Koppi, T., Inclusion and Online Learning Opportunities: Designing for Accessibility, ALT-J, vol. 10, no. 2, 2002, pp 17-28

[26] Bergström, L, Grahn, K.J. and Pulkkis,G, A Virtual Learning Environment for Mobile IP, Issues in Informing Science and Information Technology, vol. 3, 2006, pp 83-101

[27] Buzzetto-More N.A. , Navigating the Virtual Forest: How Networked digital Technologies can Foster Transgeographic Learning, Issues in Information Science and Information Technology, vol. 3, 2006, pp 104-147

[28] Murray, I (2009), eLearning Modalities and the Vision Impaired, PhD Thesis, Curtin University of Technology, Perth, Western Australia

[29] Dulin, D, 2007, Effects of the use of raised line drawings on blind people's cognition, European Journal of Special Needs Education, Volume 22, Issue 3, pp 341 - 353

[29] J. Hope, B. vonKonsky, I. Murray, L. C. Chew and B. Farrugia, A Cisco Education Tool Accessible to the Vision Impaired, ASSETS06, Portland, Oregon USA, October 23-25, 2006, pp235-236

[31] Armstrong, H and Murray, I, Remote and Local Delivery of Cisco Education for the Vision-Impaired, ITiCSE 2007, Dundee Scotland, 25-27 June 2007, pp78-81

Encouraging the Participation and Knowledge Reuse in Communities of Practice by Using a Multi-Agent Architecture and a Trust Model

Aurora Vizcaíno, Juan Pablo Soto, Javier Portillo-Rodríguez, Mario Piattini
Alarcos Research Group – Institute of Information Technologies & Systems
Dep. of Information Technologies & Systems – Escuela Superior de Informática
University of Castilla – La Mancha
Ciudad Real, Spain
{aurora.vizcaino, javier.portillo, mario.piattini}@uclm.es, jpsotob@gmail.com

Abstract — **This paper proposes a multi-agent architecture based on the concepts of communities of practice and trust to manage knowledge management systems. The main goal of this proposal is to assist community of practice members in deciding what or who to trust and in this way attempt to foster the reuse of information in organizations which use knowledge management systems. One contribution of this work is a trust model which takes into account certain factors that human beings consciously or unconsciously use when they have to decide whether or not to trust in something or somebody. Moreover, in order to illustrate how the model can be used, a prototype with which to recommend documents is also described.**

Keywords — Multi-agent System, Communities of Practice, Trust, Knowledge Management.

I. Introduction

Traditional Knowledge Management Systems (KMS) have received certain criticism as they are often implanted in companies overloading employees with extra work; for instance, employees have to introduce information into the KMS and worry about updating this information. As result of this, these systems are sometimes not greatly used since the knowledge that these systems have is often not valuable or on other occasions the knowledge sources do not provide the confidence necessary for employees to reuse the information. For this, companies create both social and technical networks in order to stimulate knowledge exchange. An essential ingredient of knowledge sharing information in organizations is that of Communities of Practice (CoPs). CoPs are becoming increasingly more common in organizations due to the fact they are a means of sharing knowledge [2] [3]. They are frequently defined as groups of people who share a concern, a set of problems, or a passion about a topic and who extend their knowledge and expertise in this area by interacting on an ongoing basis [4]. However, CoPs members are ever-increasingly distributed throughout different geographic locations. This implies a lack of face-to-face communication which affects certain aspects of interpersonal relationships. For instance, if people never experience face-to-face communication and only use groupware tools to communicate, then trust often decreases [5][33]. This lack of trust makes it more difficult for CoPs members to know which of their fellow-members are more trustworthy. This presents a problem, as in CoPs the main knowledge sources are the members themselves. We thus consider that it is highly important to be able to discover how trustworthy a knowledge source (i.e. another member) is. This knowledge will help members to decide whether or not a piece of knowledge is valuable depending on the knowledge source from which it originates. Therefore, in order to support CoPs members in this task, this paper describes a trust model designed solely for CoPs in which various psychological aspects that a person uses, either consciously or unconsciously, to value whether another person is trustworthy have been considered. This model has been used in the implementation of a prototype in which software agents make recommendations to users about what documents are most relevant to them according to their preferences and trust in knowledge sources.

In the following section we describe the multi-agent architecture proposed. Later, the next section describes the trust model that we propose. Section Four explains the details of how this model was implemented in a prototype. Section Five outlines related work. Finally, in Section Six, our conclusions are summarized.

II. A multi-agent architecture

The multi-agent architecture proposed is composed of two levels (see Figure 2): reactive and deliberative-social. The reactive level is considered by other authors to be a typical level that a Multi-Agent System (MAS) must have [6]. A deliberative level is often also considered as a typical level but a social level is not frequently considered in an explicit way, despite the fact that these systems (MAS) are composed of several individuals, the interactions between them and the plans constructed by them. The social level is only considered in those systems that attempt to simulate social behavior. Since we wish to emulate human feelings such as trust and intuition when working in CoPs, we have added a social level that considers the social aspects of a community and which takes into account the opinions and behavior of each of the members of that community. Other previous works have also added a social level. For example, Imbert & de Antonio [7] attempt to emulate human emotions such as fear, thirst or bravery, but in this case the author uses an architecture made up of three levels: reactive, deliberative and social. In our case the deliberative and social levels are not separate levels since we have realized that plans created in the deliberative level involved social interactions. We therefore consider that, in our case, it might be more efficient to define a level which is composed of two parts (deliberative-social level) rather than considering two separate levels.

Each of these levels is explained in greater detail in the following subsections.

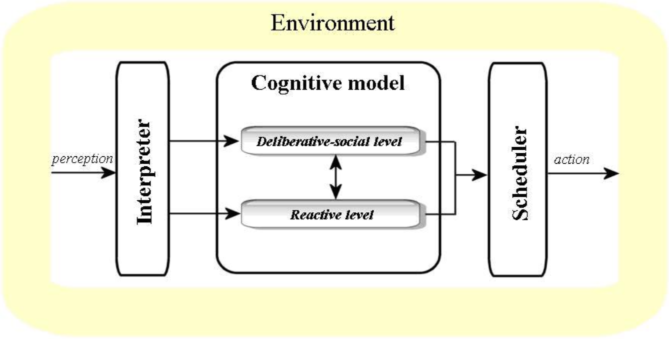


Figure 1. Multi-agent architecture

Two further important components of our architecture are the Interpreter and the Scheduler. The former is used to perceive the changes that take place. The Scheduler indicates how the actions should be executed.

2.1 Reactive level

This is the level in charge of perceiving changes in its environment and respond to these changes at the precise moment at which they happen, for instance when an agent will execute another agent's request without any type of reasoning. The components of the reactive level are (see Figure 2):

Internal model. This component stores the individual's features. These features will be consulted by other agents in order to discover more about the person represented by the User Agent. In the case of CoPs, the members will be also knowledge sources since they contribute to the CoP with information. Therefore, the model stores the following information, which will be useful in calculating how trustworthy a knowledge source is:

- *Expertise*. This information is an important factor since people often trust experts more than novice employees. The level of expertise that an individual has in a CoP could, for example, be calculated, from his/her CV or by considering the amount of time that a person has been working on a topic.
- *Position*. Employees often consider information that comes from a superior as being more reliable than that which comes from another employee in the same (or a lower) position as him/her [8]. However, this is not a universal truth and depends on the situation. For instance, in a collaborative learning setting collaboration is more likely to occur between people of a similar status than between a superior and his/her employee or between a teacher and pupils [9]. Such different positions inevitably influence the way in which knowledge is acquired, diffused and eventually transformed within the CoP.
- *Profile*. This part is included in the internal model to describe the profile of the person that the agent is acting on behalf of. Therefore, a person's preferences are stored here.

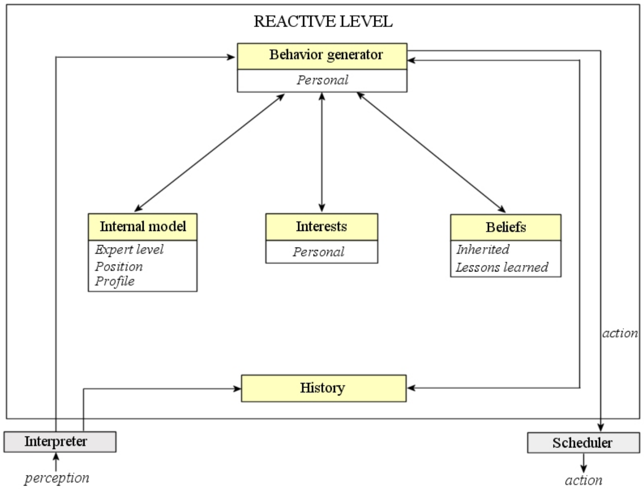


Figure 2. Reactive level

Beliefs. This module is composed of inherited beliefs and lessons learned from the agent itself. Inherited beliefs are the organization's beliefs that the agent receives such as the enterprise's organizational diagram or the organization's philosophy. Lessons learned are the lessons that the agent obtains while it interacts with the environment.

Interests. They are a special kind of beliefs. This component represents individual interests that an agent has about a topic or about a knowledge source.

Behavior generator. This component is fundamental to our architecture. It is here that the actions to be executed by the agent are triggered. To do this, the behavior generator considers various information which comes from the internal model or the agent's interests and beliefs. This information is used by the behavior generator to generate an action, such as answering question about the level of expertise that the person who the agent represents has.

History. The history component stores the agent's interactions with its environment. This information represents the received by the interpreter and stored in the agent history. The history component also registers each of the actions executed by the agent in the environment. Finally, all the information stored by this component can be used to discover the knowledge sources which are most frequently consulted by or useful to the agents in the community.

2.2 Deliberative-social level

At this level, the agent has a type of behaviour which is oriented towards objectives, that is, it takes the initiative in order to plan its performance with the purpose of attaining its goals.

The components of the deliberative-social level are (see Figure 3):

Goals generator. Depending on the state of the agent, this module must decide what the most important goal to be achieved is.

Social beliefs. This component represents a view that the agent has of the communities and their members. For instance, beliefs about other agents.

Social interests. This is a special type of belief. In this case it is represented interest about other agents.

Intuitions. As we are modelling community members we have attempted to introduce factors into this architecture that influence people when they need to make decisions about whether or not to trust a knowledge source. One of these factors is intuition, which is a subjective factor since it depends on the individual person. This concept is highly important when people do not have any previous experience. Other authors have called this issue "indirect reputation or prior-derived reputation" [10]. In human societies, each of us probably has different prior beliefs about the trustworthiness of strangers we meet. Sexual or racial discrimination might be a consequence of such prior belief [10]. We often trust more in people who have similar features to our own. For instance, when a person consults a community for rating products or services such as *Tripadvisor* [11], s/he often checks comments from people who are of the same age or have similar interests to him/her. In this research, intuition has therefore been modeled according to the similarity between agents' profiles: the greater the similarity between one agent and another, the greater the level of trust. The agents' profiles may change according to the community in which they are working.

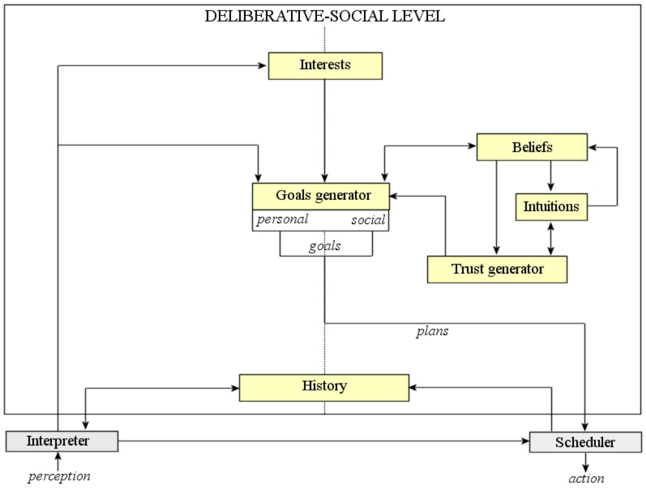


Figure 3. Deliberative-social level

Trust generator. This module is in charge of generating a trust value for the knowledge sources with which an agent interacts in the community. To do this, the trust generator module considers the trust model explained in detail in [12] which considers the information obtained from the internal model and the agent's intuitions.

III. THE TRUST MODEL

It is first important to clarify that this trust model was designed to be used in companies in which CoPs are created as a knowledge management strategy with the goal of sharing knowledge and reusing lessons learnt. The word 'employees' therefore appears in this paper on several occasions, as it is assumed that the final aim of this research is to support companies, enterprises and organizations in general in the creation and use of CoPs as a means of improving their knowledge management.

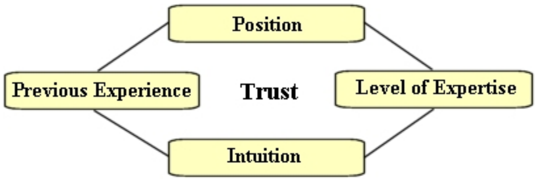


Figure 4. Trust factors

There are many recent proposals for reputation mechanisms and approaches to evaluate trust in P2P systems in general [13, 14], and multi-agent systems in particular [15, 16, 17, 18, 19, 20, 21]. However, there is no universal agreement on the definition of the trust and reputation. Since the main goal of our work is to rate the credibility of information sources and of knowledge in CoPs, it is first necessary to define these two important concepts.

Trust is a complex notion whose study is usually of a narrow scope. This has given rise to an evident lack of coherence among researchers in the definition of trust. For instance Wang & Vassileva define trust as a peer's belief in another peer's capabilities, honesty and reliability based on his/her own direct experiences [13].

Another important concept related to trust is reputation. Several definitions of reputation can be found in literature, such as that of Barber & Kim whom define this concept as the amount of trust that an agent has in an information sources [18], created through interactions with information sources, and that of Mui *et al* [22] which define reputation as a perception a partner creates through past actions about his intentions and norms. This may be considered as a global or personalized quantity [22].

These concepts of trust and reputation are sometimes used interchangeably. However, recent research has shown that there is a clear difference between them, whilst accepting that there is a certain amount of correlation between the two concepts in some cases [23, 24].

In our work we intend to follow the definition given by Wang & Vassileva [13] which considers that the difference between both concepts depends on who has previous experience, so if a person has direct experiences of, for instance, a knowledge source we can say that this person has a trust value in that knowledge.

Many authors consider that trust facilitates problem solving by encouraging information exchange [25]. However, the development of trust in a virtual setting is often more difficult than in co-located meetings [26]. Moreover, the idea of trusting or not trusting in something or somebody is context dependent. For instance, at an auction people may attempt to cheat in order to obtain greater benefits. Furthermore, in a CoP other factors may arise which might be objective and sub-objective. Both types have been considered in this model (see Figure 4), since both are frequently relevant in the personal decision-making processes.

The first is that of the **Position** that a person holds in the organization in which the CoPs exist. This factor will be calculated in our research by considering a weight that can strengthen this factor to a greater or to a lesser degree. This is an objective factor since it is provided or indicated by an exterior entity (for instance, it may be provided by the organization, by the community itself, etc).

Level of Expertise (**LE**): this term can be briefly defined as the skill or knowledge of a person who knows a great deal about a specific thing. This is an important factor since people often trust in experts more than in novice employees. In addition, an "individual" level of knowledge is embedded in the skills and competencies of the researchers, experts, and professionals working in the organization [27].

This factor can be seen as objective or sub-objective according to where this concept originates. For instance if it is specified by the organization it will be considered as objective. However, if its value is provided by the opinion of another agent then it will be seen as a sub-objective value.

Previous experience (**PE**): A trusting decision is based on the truster's relevant prior experiences and knowledge [28, 29]. Experiences and knowledge form the basis of trust in future familiar situations [30]. Consequently, members of CoPs have greater trust in those knowledge sources from which they have previously obtained more "valuable information". Therefore, previous experience increases or decreases trust, and this factor can be very useful in detecting trustworthy knowledge sources in CoPs. In this case this factor is subjective since it depends on a person's opinion.

Intuition (**I**): When people do not have any previous experience they often use their "intuition" to decide whether or not they are going to trust something. In this research, intuition has been modelled according to the similarity between agents' profiles: the greater the similarity between one agent and another, the greater the level of trust. This is, of course, a highly subjective value because it is almost at the same level as a hunch and depends directly on the point of view of each person.

As will later be explained, it is possible to decide to place more importance upon one factor or another according to the setting in which the trust model is used. For this reason, we have pondered each factor with a weight which emphasizes a factor or decreases its importance. An explanation of how to use this model will be shown in the following section.

IV. A Prototype to recommend documents

In order to test the trust model, a prototype with which to recommend documents to CoP members was developed. This prototype allows CoP members to introduce documents relating to different topics. Each time a person uses a document recommended by this tool, that person should evaluate it to enable the prototype to obtain user-feedback.

The prototype was developed by using the software architecture described in section 2, This section will centre on explaining how agents calculate each factor of the trust model explained in the previous section, and which is considered in the following formula:

$$T_{\boldsymbol{ij}} = wp*P_j + we*LE_j + wi*I_{ij} + PE_{ij} \qquad (1)$$

Let us then imagine that an agent *i* must evaluate how trustworthy another agent *j* is. It will therefore use Formula (1) in which $T_{\boldsymbol{ij}}$ is the value of *j*'s trust in the eyes of *i*. We shall now describe how each factor of the formula is calculated.

Position

When a new member joins a community that person must indicate his/her position within the organization and his/her software agent will calculate the Position (P) value of that person by using the following formula:

$$P = UPL/NL \qquad (2)$$

where UPL is the user's position level and NL is the number of levels in the community.

Therefore, if a community, for instance, has 5 possible position levels then NL=5, and if the new member has a level of UPL=2 then the value of P will be 2/5=0.4. Therefore, the different values of P for a community with five levels will be those shown in Table 1:

TABLE I. Example of position levels

Levels	Values P
1	0.2
2	0.4
3	0.6
4	0.8
5	1

The P values will always be between 0 and 1. Moreover, situations may exist in which P will not been taken into account, for instance in those CoPs in which all the members have the same level or whose members do not wish to consider this criterion. In these cases wp (weight of position) will be zero and position will not be considered in the formula. A further situation exists in which wp is equal to zero. This occurs when the value of the PE > U (U being a threshold which is chosen when creating the community). In this case, the agent will use the following formula to calculate the wp value:

$$wp = int\ (U/PE_{ij})\ \text{being}\ PE_{ij} > 0$$

where U is a threshold of previous experience. PE_{ij} is the value of previous experience of an agent *i* with another agent *j*.

Thus, when PE_{ij} is greater than a particular threshold U, wp will be 0, thus ignoring the position factor. However, when one agent does not have enough PE of another it may use other factors to obtain a trust value. On the other hand, when the agent has had a considerable amount of PE with this agent or with the knowledge that it has provided then it is more appropriate to give more weight to this factor, since PE is the key factor in all trust models, as will be described in Section 4. Therefore, if an agent *j* has a high value of position but most of agent *i*'s previous experience of *j* has not been successful then the position will be ignored. This thus avoids the situation of, for instance, a boss who does not contribute with valuable documents but is considered trustworthy solely because s/he is a boss.

Level of Expertise

As was previously mentioned, this factor is used to represent the level of knowledge and know-how that a person has in a particular domain. In this prototype this factor may change since a person may become more expert in a topic as time goes by.

In this tool, when creating a community the levels of expertise considered is also indicated, for instance: novice, beginner, competent, expert and master. Each time a new member joins a community s/he will indicate the level of expertise that s/he considers him/herself to have. If the members of the community and their level of expertise are known to the creator of the community then that person can introduce them in the tool. Once the level of expertise has been introduced, the user agent will calculate the value for this level by using the following formula:

$$LE_j = L_j/NT+AV_j \qquad (3)$$

where L_j is the level of expertise that was introduced, and NT is the number of levels in the community. The term AV_j is the Adjustment Value for agent *j*. This term is extremely important since it will be used to adjust the experience of each user. This term was introduced with the goal of avoiding two situations:

- That a person either deliberately or mistakenly introduces a level of experience that is not the level that s/he has.
- That, whilst in the community, a person becomes more expert leading to the situation that his/her level of expertise should be adjusted.

Initially AV_j will be 0, and each time a member interacts with a document or information provided by *j* the member will rate this document or information and send this evaluation to an agent called the manager agent which is in charge of managing the community. The manager agent will verify whether the evaluation is negative or positive. If it is positive, then agent *j*'s level of experience can be modified by calculating AV_j as:

$$AV_j = (VL_n - VL_{n-1})/PT \quad (n \neq 1)$$

If it is negative, then:

$$AV_j = -\ (VL_n - VL_{n-1})/PT \quad (n \neq 1)$$

where VL_n is the value that a particular level of experience has. PT is the Promotion Threshold which is used to determine the number of positive rates necessary to promote a superior level of experience. Let us illustrate this with an example. In a community there are four levels with the following values.

TABLE II. POSITION VALUES

Labels	Level(n)	Value(VL)
Novice	1	0
Beginner	2	0.25
Competent	3	0.5
Expert	4	0.75
Master	5	1

In this case, the difference between the levels is 0.25 as:

$$VL_n - VL_{n-1} = 0.25.$$

In this version of the tool it is assumed that at least 5 rates are necessary to change the level so PT will be 5, and AV_j will be 0.25/5=0.05. This is therefore the value that will be added when a positive rate is received or that will be subtracted when this rate is negative. With five positive rates (5*0.05=0.25) there is thus a level promotion.

Intuition

This factor is used when the Previous Experience is low and it is necessary to use other factors to calculate a trust value. This is one contribution of our work, since most of the earlier trust models are based solely on previous experience. The agents compare their own profiles with the other agents' profiles in order to decide whether a person appears to be trustworthy or not. Therefore, the more similar the profiles of two agents are, for instance *i* and *j*, the greater the I_{ij} value in formula (1) will be. We could say that an agent 'thinks' "I do not know whether I can trust this agent but it has similar features to me so it seems trustworthy". The agents' profiles may alter according to the community in which they are working. In our case, as the data stored in the agents' profiles are 'position' and 'expertise', both these features will be taken into account. Therefore, the factors that the tool compares are:

- Experience Difference (ED)
- Position Difference (PD)

Thus, the Intuition value of an agent *i* about *j* (I_{ij}) is:

$$I_{ij} = ED_{ij} + PD_{ij} \qquad (4)$$

where $ED_{ij} = LE_i\text{-}LE_j$ and $PD_{ij} = P_i\text{-}P_j$

This formula (4) is based on the idea that a person normally has a greater level of trust in people who have a higher level of experience or who are in a higher position than that person him/herself. Hence, when an agent compares its profile with another agent with higher values, the value of intuition will be positive. Let us consider the case of agent *i* which has values of LE_i=0.75 and P_i=0.25. This agent wishes to know how trustworthy another agent *j* is. In this case the agent will use Formula (1) and, depending on the information that it has about *j*, it will or will not be necessary for it to calculate the intuition factor. In this situation we shall suppose that there is little previous experience and that this must be calculated. The values for the agent *j* are LE_j=0.25 and P_j=0.5. As Figure 2 shows:

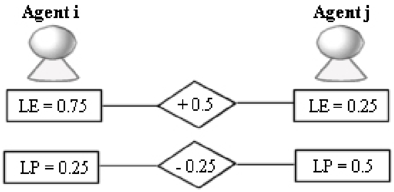


Figure 5. Comparing profiles

I_{ij}=0.25 as ED_{ij}=0.5 and $PD_{ij=}$-0.25

As with position, intuition will or will not be calculated depending on the level of PE. Thus, the weight of intuition, (see Formula 1) *wi* will be calculated as follows:
wi = int (U/PE_{ij}) with $PE_{ij} \neq 0$.

Previous experience

This factor is the most decisive of all the factors in Formula (1). In fact, all the previous factors depend on it as an agent will decide whether or not to use the remaining factors according to the value of Previous Experience (PE). Previous Experience is obtained through the interactions that the agent itself has, so this is direct experience. Each time one agent interacts with another (by interacting we mean that one agent uses a document provided by another), the first agent asks its user to rate that document in order to discover whether the document was: useful for him/her, related to the topic at hand, recommendable for other people interested in the same topic, up-to-date.

TABLE III. PE LABELS

Label	PE Level
Very Bad	- 0.3
Bad	- 0.2
Medium	+ 0.1
Good	+ 0.2
Very good	+ 0.3

The agent then labels this interaction with a label from Table 3. A value for Current Experience (CE) is thus obtained which will modify the previous value of PE in accordance with the following formula:

$$PE_{ij}(x) = PE_{ij}(x\text{-}1) + CE_{ij}(x) \qquad (5)$$

where $PE_{ij}(x)$ is the value of Previous Experience that the agent *i* has about another agent *j* in an interaction x.

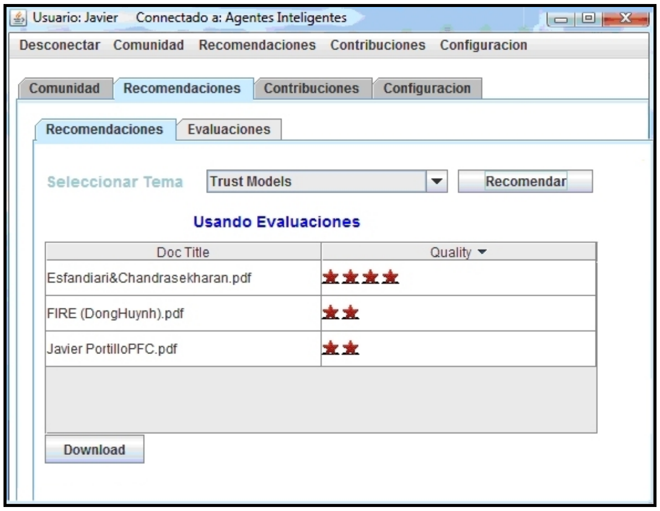
$EP_{ij}(x\text{-}1)$ is the value of Previous Experience that the agent *i* had about another agent *j* before the interaction x.

$CE_{ij}(x)$ is the value of the experience that *i* has had with *j* in the interaction x.

For instance, if an agent *i* has just taken part in an interaction with the agent *j*, and this is labeled as "bad", but the value of $PE_{ij}(x\text{-}1)$ was 0.8, then the value of $PE_{ij}(x)$ will be 0.6 obtained from (0.8+(-0.2)). Moreover the agent *i* will send the manager agent the value of $CE_{ij}(x)$ in order to calculate AV_j (see Level of Expertise).

As has previously been explained, the Position and Intuition factors depend on the PE value. When an agent has sufficient PE then Position and Intuition can be ignored, and only the PE and the LE will be considered. The latter is also included to ensure that an agent takes advantage not only of its own previous experience but also of that of the other agents since Level of Expertise (LE) is adjusted by the AV_j which comes from other previous experience.

In order to illustrate how the prototype works, let us look at an example. If a user selects a topic and wishes to search for documents related to that subject, his/her user agent will contact other user agents which have documents related to the theme at hand. The user agent will then calculate the trust value for each agent, meaning that these agents are considered to be knowledge sources and the user agent needs to calculate which "knowledge source" is more trustworthy. Once these values have been calculated, the user agent shows its user only the documents which have come from the most trustworthy agents.



Seleccionar Tema: Trust Models

Usando Evaluaciones

Doc Title	Quality ▼
Esfandiari&Chandrasekharan.pdf	★★★★
FIRE (DongHuynh).pdf	★★
Javier PortilloPFC.pdf	★★

Figure 6. List of documents recommended

Figure 6 shows the results that the User Agent would display after the documents had been sorted by the trust value obtained.

V. RELATED WORK

This research can be compared with other proposals that use agents and trust models in knowledge exchange. Caballero *et al* [19] present a trust and reputation model that considers trust and reputation as emergent properties of direct interactions between agents, based on multiple interactions between two parties. In this model, trust is a belief an agent has about the performance of the other party to solve a given task, according to own knowledge. Abdul-Rahman & Hailes propose a model which allows agents to decide which agents' opinions they trust more and to propose a protocol based on recommendations [25]. This model is based on a reputation or word-of-mouth mechanism. The main problem with this approach is that every agent must maintain rather complex data structures which represent a kind of global knowledge about the whole network.

Barber and Kim present a multi-agent belief revision algorithm based on belief networks [18]. In their model the agent is able to evaluate incoming information, to generate a consistent knowledge base, and to avoid fraudulent information from unreliable or deceptive information sources or agents. This work has a similar goal to ours. However, the means of attaining it are different. In Barber and Kim's case reputation is defined as a probability measure, since the information source is assigned a reputation value of between 0 and 1. Moreover, every time a source sends knowledge, that source should indicate the certainty factor that the source has of that knowledge. In our case, the focus is very different since it is the receiver who evaluates the relevance of a piece of knowledge rather than the provider as in Barber and Kim's proposal.

Huynh *et al* [15] present a trust and reputation model which integrates a number of information sources in order to produce a comprehensive assessment of an agent's likely performance. In this case the model uses four parameters to calculate trust values: interaction trust, role-based trust, witness reputation and certified reputation. We use certified reputation when an agent wishes to join a new community and uses a trust value obtained in other communities, but in our case this certified reputation is made up of four factors and is not only a single factor.

Also, works such as Guizzardi *et al* [31] use the term 'Community' to support knowledge management but a specific trust model for communities is not used.

The main differences between these reputation models and our approach are that these models need an initial number of interactions to obtain a good reputation value and it is not possible to use them to discover whether or not a new user can be trusted. A further difference is that our approach is orientated towards collaboration between users in CoPs. Other approaches are more orientated towards competition, and most of them are tested in auctions.

VI. CONCLUSIONS

CoPs are a means of knowledge sharing. However, the knowledge reused should be valuable for the members, otherwise CoP members might prefer to ignore the documents that a community has. In order to encourage the reuse of documents in CoPs, in this work we propose a multi-agent system to suggest trustworthy documents. Some of the advantages of our system are:

- The use of agents to represent members of the community helps members to avoid the problem of information overload since the system gives the User Agents the ability to reason about the trustworthiness of the other agents or about the recommendation of the most suitable documents to the members of the community. Users are not, therefore, flooded with all the documents that exist with regard to a particular topic, but their User Agents filter them and only recommend the most trustworthy or those which are provided by more trustworthy sources or sources which have preferences and features that are similar to them.
- Detecting whether members store documents that are not useful, since the system provides users with the opportunity to evaluate the documents consulted, and when a document is frequently evaluated with low marks then the Manager Agent will check who the provider is and whether most of that person's documents have a low evaluation. In this case, two options can be considered. First that the person does not have enough knowledge about the topic, in which case the Manager Agent can consult the Level of Expertise that this person has (which is indicated when a person

joins a community), and if this level is not suitable the Manager Agent can modify it. The second option is that this person may be consciously introducing invaluable documents. In this case the trust in this source will be low and the documents will rarely be recommended. The system can also detect the users with the greatest participation and those whose documents have obtained higher rates. This information can be used for two purposes: expert detection and/or recognition of fraudulent members who contribute with worthless documents. Both functionalities imply several advantages for any kind of organization; for instance, the former permits the identification of employee expertise and measures the quality of their contributions, and the latter permits the detection of fraud when users contribute with non-valuable information.

- The system facilitates the exchange and reuse of information, since the most suitable documents are recommended. Furthermore, this tool can be understood as a knowledge flow enabler [32], which encourage knowledge reuse in companies.

On the other hand thanks to the trust model the agents can calculate a trust value even though the community has only recently been created since, in order to calculate trust, various known factors are used such as Position, Level of Expertise and even Profile Similarity. This is a key difference with regard to other models which use only previous experience and which cannot then calculate trust values if the system is just starting to work. When a new member arrives it is also impossible for other models to calculate a previous trust value related to this new member. Moreover, the model helps to detect an increasing problem in companies or communities in which employees are rewarded if they contribute with knowledge in the community. Thus, if a person introduces, for instance, non-valuable documents with the sole aim of obtaining rewards, the situation can be detected since these documents will have lost trust values and the person will also considered to be less trustworthy. The agent will, therefore, not recommend those documents. Moreover, the formulas proposed are very simple and easy to understand. This is an advantage aver the previous models which are often not greatly used since they are difficult to implement.

Acknowledgment

An earlier version of this articles was presented at the International Conference on Information, Process, and Knowledge Management (eKNOW'09)[1].

This work is partially supported by FABRUM project. Ministerio de Ciencia e Innovación (grant PPT-430000-2008-063), PEGASO (TIN2009-13718-C02-01), MELISA (PAC08-0142-3315) and ENGLOBAS (PII2109-0147-8235) projects, Junta de Comunidades de Castilla-La Mancha, Consejería de Educación y Ciencia, in Spain, and CONACYT (México) under grant of the scholarship 206147 provided to the second author.

References

1. A. Vizcaíno, J. Portillo-Rodríguez, Soto, J.P., M. Piattini, "Encouraging the Reuse of Knowledge in Communities of Practice by Using a Trust Model", in International Conference on Information Process, and Knowledge Management (eKNOW), IEEE Computer Society, pp. 28-33, Cancun México, 2009.
2. Y. Malhotra, Knowledge Management and Virtual Organizations, IGI Publishing, Hershey, PA, USA. 2000.
3. H. Gebert, M. Geib., L. Kolbe and Riempp, G. "Knowledge-enabled Customer Relationship Management - Integrating Customer Relationship Management and Knowledge Management Concepts", in Journal of Knowledge Management, vol. 7(5), pp. 107-123, 2003.
4. E. Wenger, R. McDermott and W. Snyder, "Cultivating Communities of Practice", in Harvard Business School Press, 2002.
5. P. Hinds and C. McGrath, "Structures that work: social structure, work structure and coordination ease in geographically distributed teams", in 20th Anniversary Conference on Computer Supported Cooperative Work. 2006. Banff, Alberta, Canada.
6. H. Ushida, Y. Hirayama and H. Nakajima, "Emotion Model for Life like Agent and its Evaluation", in Proceedings of the Fifteenth National Conference on Artificial Intelligence and Tenth Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI), 1998.
7. R. Imbert, and A. de Antonio, "When emotion does not mean loss of control" in T. Panayiotopoulos, J. Gratch, R. Aylett, D. Ballin, P. Olivier, and T. Rist (Eds), LNCS, Springer-Verlag, pp. 152-165, 2005.
8. S. Wasserman and J. Glaskiewics, "Advances in Social Networks Analysis", Sage Publications, 1994.
9. P. Dillenbourg, "Introduction: What Do You Mean By 'Collaborative Learning'?", in Collaborative Learning Cognitive and Computational Approaches. Dillenbourg (Ed.). Elsevier Science., 1999.
10. L. Mui, A. Halberstadt and M. Mohtashemi, "Notions of Reputation in Multi-Agents Systems: A Review", in International Conference on Autonomous Agents and Multi-Agents Systems (AAMAS), pp. 280-287, 2002.
11. http://www.tripadvisor.com
12. J.P. Soto, A. Vizcaíno, J. Portillo and M. Piattini, "Applying Trust, Reputation and Intuition Aspects to Support Virtual Communities of Practice", in 11th International Conference on Knowledge-Based and Intelligence Information and Engineering Systems (KES), LNCS 4693, Springer-Verlag, pp. 353-360, 2007.
13. Y. Wang and J. Vassileva, "Trust and Reputation Model in Peer-to-Peer Networks", in Proceedings of the 3rd International Conference on Peer-to-Peer Computing, pp. 150-157, 2003.
14. B. Yu, M. Singh and K. Sycara, "An evidential model of distributed reputation management", in Proceedings of the first international joint conference on Autonomous Agents and Multiagents Systems (AAMAS), New York, USA: ACM Press, pp. 294-301, 2002.

15. T. Huynh, N. Jennings and N. Shadbolt, "FIRE: an integrated trust and reputation model for open multi-agent systems", in Proceedings of 16th European Conference on Artificial Intelligence, pp. 18-22, 2004.

16. J. Sabater and C. Sierra, “Reputation and Social Network Analysis in Multi-Agent Systems”, in proceedings of the first international joint conference on autonomous agents and multiagent systems (AAMAS), New York, USA: ACM Press, pp. 475-482, 2002.

17. W. Taecy, G. Chalkiadakis, A. Rogers and N. Jennings, “Sequential Decision Making with Untrustworthy Service Providers”, in Proceedings of 7th International Conference on autonomous Agents and Multiagent Systems (AAMAS), pp. 755-762, 2008.

18. K. Barber and J. Kim, "Belief Revision Process Based on Trust: Simulation Experiments", in 4th Workshop on Deception, Fraud and Trust in Agent Societies, pp. 1-12, 2004.

19. A. Caballero, J. Botia and A. Skarmeta, "A New Model for Trust and Reputation Management with an Ontology Based Approach for Similarity Between Tasks", in Fischer, K., Timm, I., André, E., Zhong, N. (eds), LNCS 4196, pp. 172-183, 2006.

20. R. Hermoso, H. Billhardt and S. Ossowski, “Integrating trust in virtual organizations”, in International Workshop on Coordination, Organization, Institutions and Norms in Agents Systems II: AAMAS nad ECAI, COIN 2006, LNAI, pp. 19-31, 2007.

21. J. Carter, E. Bitting, and A. Ghorbani, “Reputation for an information-sharing multi-agent system”, in Computational Intelligence, vol. 18, no. 2, pp. 515-534, 2002.

22. L. Mui, M. Mohtashemi, and A. Halberstadt, “A Computational Model of Trust and Reputation for E-businesses”, in Proceedings of the 35th Hawaii International Conference on Systems Sciences (HICSS), IEEE Computer Society Press, 2002.

23. A. Jøsang, R. Ismail, and C. Boyd, “A Survey of Trust and Reputation Systems for Online Services Provision”, in Decision Support Systems, vol. 43, no. 2, pp. 618-644, 2007.

24. J. Sabater and C. Sierra, “*Review on Computational Trust and Reputation Models*”, in Artificial Intelligence Review, vol. 24, pp. 33-60, 2005.

25. A. Abdul-Rahman and S. Hailes, "Supporting Trust in Virtual Communities", in Proceedings of 33rd Hawaii International Conference on Systems Sciences (HICSS'00), IEEE Computer Society, vol. 6, pp. 1769-1777, 2000.

26. B. Misztal, "Trust in Modern Societies", Polity Press, Cambridge MA, 1996.

27. I. Nonaka and H. Takeuchi, "The Knowledge Creation Company: How Japanese Companies Create the Dynamics of Innovation", Oxford University Press, 1995.

28. R. Hardin, "The Street Level Epistemology of Trust", in Politics and Society, vol. 21, pp. 505-531, 1993.

29. A. Jøsang, "The Right Type of Trust fro Distributed Systems", in New Security Paradigms, 1996.

30. N. Luhmann, "Trust and Power", in Wiley, Chichester, 1979.

31. R. Guizzardi-Silva, L. Aroyo and G. Wagner, "Help&Learn: A Peer-to-Peer Architecture to Support Knowledge Management in Collaborative Learning Communities" in Revista Brasileira de Informatica na Educação, vol. 12(1), pp.29-36, 2003.

32. O. Rodríguez-Elias, A. Martínez-García, A. Vizcaíno, J. Favela and M. Piattini, "A Framework to Analyze Information Systems as Knowledge Flow Facilitators", Information Software Technology, vol. 50(6), pp. 481-498, 2007.

33. J.P. Soto, A. Vizcaíno, J. Portillo-Rodríguez, and M. Piattini, “Why Should I Trust in a Virtual Community Member?”, in 15th Collaboration Researchers’ International Workshop on Groupware (CRIWG), LNCS 5784, pp. 126-133, 2009.

An Ontology Learning Framework Using Focused Crawler and Text Mining

Hiep Phuc Luong
CSCE Department
University of Arkansas
Fayetteville, AR, USA
hluong@uark.edu

Susan Gauch
CSCE Department
University of Arkansas
Fayetteville, AR, USA
sgauch@uark.edu

Qiang Wang
CSCE Department
University of Arkansas
Fayetteville, AR, USA
qxw002@uark.edu

Anne Maglia
Missouri University of
Science and Technology
Rolla, MO, USA
magliaa@mst.edu

Abstract— **Manual ontology construction is costly, time-consuming, error-prone and inflexible to change. To address these problems, researchers hope that an automated process will result in faster and better ontology construction and enrichment. Ontology learning has become recently a major area of research whose goal is to facilitate the construction of ontologies by decreasing the amount of effort required to produce an ontology for a new domain. However, most of current approaches are dealing with some specific tasks or a part of the ontology learning process rather than providing complete support to users. There are few studies that attempt to automate the entire ontology learning process from the collection of domain-specific literature, filtering out documents irrelevant to the domain, to text mining to build new ontologies or enrich existing ones.**

In this paper, we present a complete framework for ontology learning that enables us to retrieve documents from the Web using focused crawling and then use a SVM (Support Vector Machine) classifier to identify domain-specific documents and perform text mining in order to extract useful information for the ontology enrichment process. Our experimental results of this framework in the amphibian morphology domain support our belief that we can use SVM and text mining approaches to improve the identification of documents and relevant words suitable for the ontology enrichment. This paper reports on the overall system architecture and our initial experiments of all phases in our ontology learning framework, i.e., document focused crawling, document classification and information extraction using text mining techniques to enrich the domain ontology.

Keywords – ontology learning; focused crawler; SVM; text mining; amphibian ontology

I. INTRODUCTION

The next generation of the Semantic Web focuses on supporting a better cooperation between humans and machines [3]. In this approach, ontologies play an important role as a backbone for providing and accessing knowledge sources. However, creating ontologies for the many and varied domains on the Web is a time-consuming process and their construction is a major bottleneck to the wider deployment and use of semantic information on the Web. Since manual ontology construction is costly, time-consuming, error-prone and inflexible to change, it is hoped that an automated process will result in a better ontology construction and create ontologies that better match a specific application [17]. These ontology learning approaches can be distinguished by the type of input used for learning, e.g., they can learn from text, from a dictionary, from a knowledge base, from a semi-structured schemata, or from a relational schemata [10] [21]. Currently, few projects attempt to support the entire ontology learning process including automated support for tasks such as retrieving documents, classifying, filtering and extracting relevant information for the ontology enrichment.

Most existing approaches for ontology learning require a large number of input documents for accurate results [20]. With the enormous growth of the Web, it is important to develop document discovery mechanisms based on intelligent techniques such as focused crawling [5] to make this process easier for a new domain. Focused crawlers go a step further than classic crawlers in order to be able to quickly collect Web pages about a particular topic or domain of the Web [8]. In our work, we use focused crawling to retrieve documents and information in a biological domain, i.e., amphibian, anatomy and morphology, by using a combination of general search engines, scholarly search engines, and online digital libraries. Due to the huge number of retrieved documents, we require an automatic mechanism rather than domain experts in order to separate out the documents that are truly relevant to the biological domain of interest. Since SVM has been recognized as one of the most successful current classification methods, we have adopted it for the classification task [23].

We have previously reported our results on collecting potential documents by using web focused crawlers, then filtering and classifying them to identify the best candidates for analysis [1]. To summarize, we found that SVM can be used to improve the identification of documents suitable for the ontology learning process. This paper extends that work in two directions. First, we present results for the information extraction process that allows us to extract the relevant information for ontology enrichment. Second, this paper describes our complete ontology learning approach and continuing work on the progress of enriching relevant vocabularies for the amphibian morphology ontology from the retrieved documents by using text mining techniques. Overall, our classification of relevant documents achieved the good prediction accuracy of 77.5% with the best-performing method of SVM algorithm (i.e., feature selection with frequency difference only). The text mining algorithm also produced good accuracy, over than 81% for all cases and reached the precision is 88% in the best case.

The goal of this research study is to implement and validate an ontology learning framework process through web focused crawling and information extraction applied to the domain of amphibian anatomy and morphology. The potential documents in this domain are gathered, classified to identify the best candidates for analysis, and then mined to extract the relevant information for the ontology enrichment process. In section 2, we present a survey of current research on ontology learning, focused crawlers, document classification, information extraction and text mining methods. In section 3, we present our ontology learning framework and its main architectural components. We also underline the process of document classifying and filtering by using SVM technique as well as the information extraction using text mining. Section 4 presents some initial experimental results for our approach. Next, we discuss on the results achieved and the usability of our work in the section 5. The final sections present our conclusions and discuss our future work in this area.

II. RELATED WORK

An ontology is an explicit, formal specification of a shared conceptualization of a domain of interest [11], where formal implies that the ontology should be machine-readable and the domain can be any that is shared by a group or community. Much of current research into ontologies focuses on construction and updating issues. In our view, there are two main approaches to ontology building: (i) manual construction of an ontology from scratch, and (ii) semi-automatic construction using tools or software with human intervention. It is hoped that semi-automatic generation of ontologies will substantially decrease the amount of human effort required in the process [13][20].

Ontology learning has recently been studied as an effective approach to facilitate the semi-automatic construction of ontologies by ontology engineers or domain experts. Ontology learning uses methods from a diverse spectrum of fields such as machine learning, knowledge acquisition, natural language processing, information retrieval, artificial intelligence, reasoning, and database management [21]. Gómez-Pérez et al. [10] present a good summary of several ontology learning projects that are concerned with knowledge acquisition from a variety of sources such as text documents, dictionaries, knowledge bases, relation schemas, semi-structured data, etc. Many of these existing approaches employ ontology learning from text documents [4], although only a few deal with ontology enrichment from documents collected from the Web. Omelayenko [20] has discusses the applicability of machine learning algorithms to learning of ontologies from Web documents and also surveys the current ontology learning and other closely related approaches. Similar to our approach, authors in [17] introduces an ontology learning framework for the Semantic Web which proceeds through ontology import, extraction, pruning, refinement, and evaluation giving the ontology engineers a wealth of coordinated tools for ontology modeling. In addition to a general framework and architecture, they have implemented Text-To-Onto system supporting ontology learning from free text, from dictionaries, or from legacy ontologies. However, they do not mention any automated support to collect the domain documents from the Web or how to automatically identify domain-relevant documents needed by the ontology learning process. Maedche et al. have presented in another paper [18] a comprehensive approach for bootstrapping an ontology-based information extraction system with the help of machine learning. They also presented an ontology learning framework which is one important step in their overall bootstrapping approach but it has still been described as a theoretic model and did not deal with the specific techniques used in their learning framework.

In another approach similar to ours, [2] has presents an automatic method to enrich very large ontologies, e.g., WordNet, that uses documents retrieved from the Web. However, in their approach, the query strategy is not entirely satisfactory in retrieving relevant documents which affects the quality and performance of the topic signatures and clusters. Moreover, they do not apply any filtering techniques to verify that the retrieved documents are truly on-topic. Inspiring the idea of using WordNet to enrich vocabulary for ontology domain, we have presented the lexical expansion from WordNet approach [15] providing a method of accurately extract new vocabulary for an ontology for any domain covered by WordNet.

Many ontology learning approaches require a large collection of input documents in order to enrich the existing ontology [20]. A common way to get these documents from the Web is to use general purpose crawlers and search engines, but this approach faces problems with scalability due to the rapid growth of the Web. In contrast, focused crawlers overcome this drawback, i.e., they yield good recall as well as good precision, by restricting themselves to a limited domain [8]. Authors in [5] describe a new hypertext resource discovery system with the purpose of selectively seeking out pages that are relevant to a pre-defined set of topics. Ester et al. [8] also introduce a generic framework for focused crawling consisting of two major components: (i) specification of the user interest and measuring the resulting relevance of a given web page; and (ii) a crawling strategy. In order to improve accuracy of the learned ontologies, the documents retrieved by focused crawlers may need to be automatically filtered by using some text classification technique such as Support Vector Machines (SVM), k-Nearest Neighbors, Linear Least-Squares Fit, TF-IDF, etc. A thorough survey and comparison of such methods and their complexity is presented in [27] and the authors in [23] conclude that SVM to be most accurate for text classification and fast training. SVM [24] is a machine learning model that finds an optimal hyperplane to separate two then classifies data into one of two classes based on the side on which they are located [6] [14].

Text mining, also known as text data mining or knowledge discovery from textual databases, refers generally to the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents [13] [25]. Tan [12] presents a good survey of text mining products/applications and aligns them based on the *text refining* and *knowledge distillation* functions as well as the *intermediate form* that they adopt. One approach similar to ours has presented a supervised ontology learning system using text mining [22]. Speretta et al used WordNet [19] similarity measures to select candidate tokens in a relatively narrow space in order to enrich the ontology. Although we share the same goal, we try to find a general and efficient way to extract a broader collection of accurate candidate tokens for ontology enrichment process that would work with any ontology.

III. Ontology Learning Framework

In this section, we first present the overall architecture of our ontology learning framework. Then, each component in this framework is described in detail in the following sections.

A. *Architecture*

Figure 1 presents the architecture of our ontology learning process framework that incorporates crawling, classifying, filtering and extracting relevant information in the amphibian and morphology domain from Internet documents. The main processes are as following (see Figure 1):

- We begin with an existing small, manually-created amphibian morphology ontology [16]. This ontology is created in the project AmphibAnat[1] with the purpose of creating a standardization of anatomy particularly pressing in amphibian morphological domain. From this ontology, we automatically generate queries for each concept in the hierarchically-structured ontology.
- We use a topic-specific spider (focused crawler) to submit these queries to a variety of Web search engines (e.g., Google, Scholar Google, Yahoo) and digital libraries. The spider downloads the potentially relevant documents listed on the first page (top-ranked) results. We also provide options to customize the number of returned results, the formats of returned documents, the list of search engines that are used to query documents, etc.
- Next, we apply SVM classification to filter out documents in the search results that match the query well but which are less relevant to the domain of our amphibian ontology.
- After the above process, we have created a collection of documents relevant to amphibian morphology. These are input to an information extraction (IE) system to mine information from documents that can be used to enrich the ontology. In our previous work [1], we planned to use a combination of pattern-based extraction methods, e.g., GATE tool [7] and statistical NLP algorithms to identify attributes to enrich the ontology. This one has been used largely by several existing researches in information extraction field. However, in this paper, we present our new results achieved by using text mining methods in the information extraction phase in order to mine new relevant vocabularies from the collection of amphibian documents. We have completed several experiments with vocabulary enrichment and this work will be further discussed in following sections.

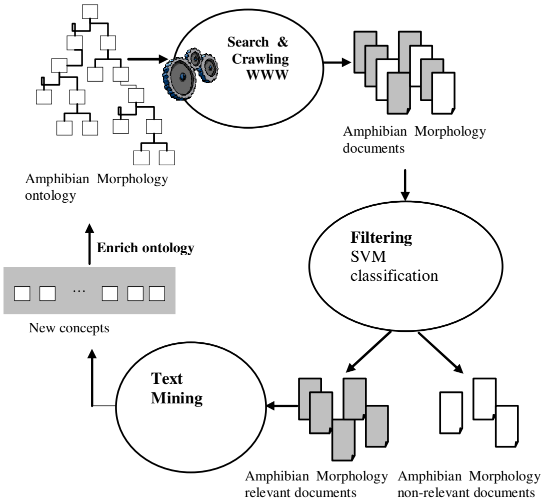

Search & Crawling WWW
Amphibian Morphology documents
Amphibian Morphology ontology
Enrich ontology
Filtering
SVM classification
...
New concepts
Text Mining
Amphibian Morphology relevant documents
Amphibian Morphology non-relevant documents

Figure 1. Architecture of ontology learning framework

B. *Amphibian Morphology Ontology*

Our proposed ontology learning framework can be used for any ontology in general domain. However, in order to validate the feasibility and effectiveness of our ontology learning approach, we have applied this framework into a specific domain, i.e., biology, anatomy and morphology, and do experiments with the Amphibian Anatomical Ontology [16].

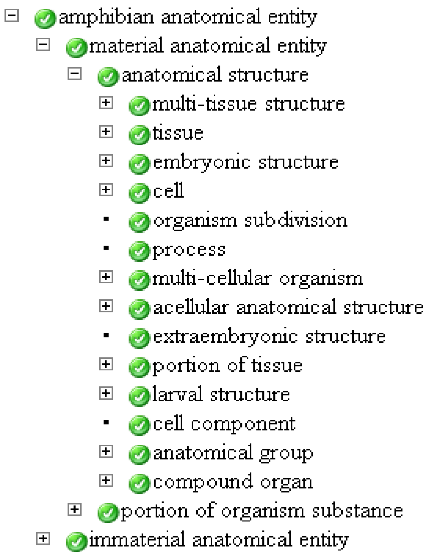
The need for terminological standardization of anatomy is particularly pressing in amphibian morphological research [16]. By standardizing the lexicon used for diverse biological studies related to anatomy, an amphibian ontology will facilitate the integration of anatomical data representing all orders of amphibians, thus enhancing knowledge representation of amphibian biology and diversity.

According to authors in [16], there are several main challenges to developing an ontology for amphibian morphology. First, the separate anatomical lexicons must be reconciled. Second, there are about 6,000 species of

[1] http://amphibanat.org/

amphibians for which the anatomical terminology must be resolved. Although much of the terminology will be similar across species, among-species variation will lead to a much larger ontology than those developed for a single model species. Third, because of anatomical diversity among amphibian orders, homologies of some structures are unknown; therefore, assigning terminological standards to them may be problematic. These challenges can be overcome if we forge a partnership between the amphibian morphological community and the power of information extraction technology. Therefore, one of the main goals of the long-term AmphibAnat (http://amphibanat.org/) NSF-sponsored project is to aim at integrating the amphibian anatomical ontology knowledge base with systematic, biodiversity, embryological and genomic resources.

Another important goal of this project is to semi-automatically construct and enrich the amphibian anatomical ontology. From a manually constructed seed ontology, we use a focused crawler and data-mining software in order to mine electronic resources for instances of concepts and properties to be added to the existing ontologies [1]. The current amphibian ontology created by this project consists of 968 different semantic concepts and 570 relationships (main properties are *is_a* and *part_of*). [16]. Figure 1 presents a part of this ontology which is available in two main formats: (i) OWL and (ii) OBO - Open Biomedical Ontology.



```
⊟ amphibian anatomical entity
    ⊟ material anatomical entity
        ⊟ anatomical structure
            ⊞ multi-tissue structure
            ⊞ tissue
            ⊞ embryonic structure
            ⊞ cell
            ▪ organism subdivision
            ▪ process
            ⊞ multi-cellular organism
            ⊞ acellular anatomical structure
            ▪ extraembryonic structure
            ⊞ portion of tissue
            ⊞ larval structure
            ▪ cell component
            ⊞ anatomical group
            ⊞ compound organ
        ⊞ portion of organism substance
    ⊞ immaterial anatomical entity
```

Figure 2. A part of the amphibian ontology

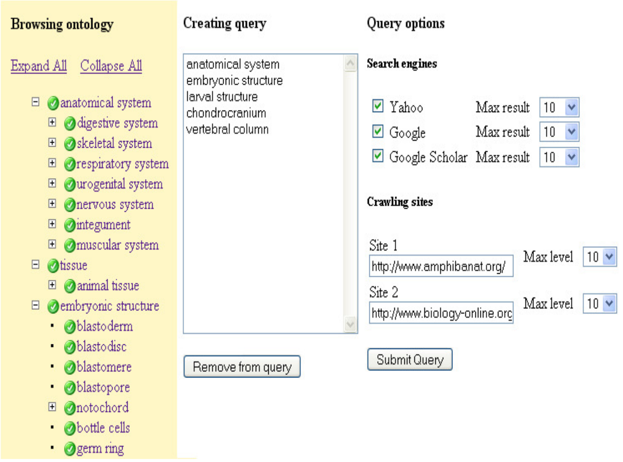
C. *Searching and Crawling Documents*

In order to collect a corpus of documents from which ontological enrichments can be mined, we use the seed ontology as input to our topic specific spider. For each concept in a selected subset of ontology, we generate a query that is then submitted to two main sources, i.e., search engines and digital libraries.

Before we could automatically generate queries from an ontology, we explored a variety of query generation strategies. To aid in this exploration, we created an interactive system that allowed us to easily create a queries and evaluate search engines. Figure 3 shows the interface to this system that enables us to create queries from existing concepts in the ontology and allows us to change parameters such as the website address, the number of returned results, the format of returned documents, etc.

From our exploration, we found that if we use the concept name, e.g., *"anatomical system"* alone as a query, we retrieve very few relevant results. However, by expanding the query containing the concept name with keywords describing the ontology domain overall, e.g., *"amphibian"* and/or *"morphology"* and also query for type of result we want, e.g., *".pdf"*, we get a larger number of relevant results. Based on these explorations, we created an automated module that, given a concept in the ontology, currently generates 3 queries with the expansion added, e.g., *"amphibian" "morphology" "pdf"*.

We next automatically submit the ontology-generated queries to multiple search engines and digital libraries related to the domain (e.g., Google, Yahoo, Google Scholar, http://www.amphibanat.org, etc.). For each query, we process the top 10 results from each search site using an HTML parser[2] to extract the hyperlinks. We have implemented some simple rules in order to automatically filter these hyperlinks to remove obviously irrelevant links, e.g., advertisement links, go-to-section links. The remaining links are then sent to the download module in order to retrieve the full documents. The results pages may contain documents in many formats, but we are interested only in HTML, pdf and text documents.



Browsing ontology

Expand All   Collapse All

```
⊟ anatomical system
    ⊞ digestive system
    ⊞ skeletal system
    ⊞ respiratory system
    ⊞ urogenital system
    ⊞ nervous system
    ⊞ integument
    ⊞ muscular system
⊟ tissue
    ⊞ animal tissue
⊟ embryonic structure
    ▪ blastoderm
    ▪ blastodisc
    ▪ blastomere
    ▪ blastopore
    ⊞ notochord
    ▪ bottle cells
    ▪ germ ring
```

Creating query

anatomical system
embryonic structure
larval structure
chondrocranium
vertebral column

Remove from query

Query options

Search engines

☑ Yahoo	Max result	10
☑ Google	Max result	10
☑ Google Scholar	Max result	10

Crawling sites

Site 1 http://www.amphibanat.org/	Max level	10
Site 2 http://www.biology-online.orc	Max level	10

Submit Query

Figure 3. Creating queries from ontology concepts for focused crawling

[2] http://htmlparser.sourceforge.net/

D. Classifying and Filtering Documents

Although documents are retrieved selectively through restricted queries and by focused crawling, we still need a mechanism to evaluate and verify the relevance of these documents to the predefined domain of amphibian morphology. We use LIBSVM classification tool [6] to separate the remaining documents into two main categories: (i) relevant and (ii) non-relevant to the domain of amphibian morphology. Only documents that are deemed truly relevant are input to the pattern extraction process.

The SVM classification algorithm must first be trained, based on labeled examples, so that it can accurately predict unknown data (i.e., testing data). The training phase consists of finding a hyperplane that separates the elements belonging to two different classes. According to [6], for median-sized problems, cross-validation might be the most reliable way to select SVM parameters so that the classifier is as accurate as possible. First, the training data is separated to several folds. Sequentially, one fold is considered as the validation set and the rest are used for training. The average of accuracy on predicting the validation sets is the cross-validation accuracy.

In our situation there are not enough examples to accurately train the classifier on all features. Thus, we may need to choose a subset of features before submitting the data to SVM [6][26]. To identify the most important features, we calculate the weights of words in documents using the KeyConcept package [9]. Each document is represented by a vector of values $wt_i * idf_i$, where wt_i is calculated by the term frequency *tf / size_of_document* (i.e., normalized by document size), and the inverse document frequency idf_i is calculated from dictionary over all documents. In section 4, we describe several feature selection methods and compare the classification results.

E. Information Extraction using Text Mining

We have so far a set of relevant documents which are closed to the domain of ontology. Our goal in this step is to extract structured and useful information from the actual text of these filtered documents. As stated in the previous section, we can use a combination of pattern-based extraction methods, e.g., GATE tool [7] and statistical NLP algorithms to identify attributes to enrich the ontology.

However, in our approach, we are aiming at producing a set of words that are most significantly related to the domain ontology by using text mining methods, then validating our algorithm. We have conducted two methods: (i) *Vector space approach* and (ii) *Part-of-speech approach* in order to calculate then rank the weights of words in relevant documents.

In the first approach, i.e., Vector space approach, we implement two algorithms, i.e., *Document-based* and *Corpus-based selection*, based on the vector space model. In order to guarantee words that are more representative of the ontology domain having higher rank values, we calculated *idf* (inverse document frequency) of words across 10,000 documents that were randomly downloaded from ODP[3] category.

*1) Document-based selection: calculates weights of words by using tf*idf*

$$W(i,j) = rtf_{(i,j)} * idf_i$$

$$rtf_{(i,j)} = \frac{tf_{(i,j)}}{N(j)}$$

$$idf_i = \log \frac{|D|}{|\{d : t_i \in d\}|}$$

with

$W(i,j)$ is the weight of term i in document j

$rtf_{(i,j)}$ is the relative term frequency of term i in document j

idf_i is the inverse document frequency of term i, which is pre-calculated across 10,000 ODP documents

$tf_{(i,j)}$ is the term frequency of term i in document j

$N(j)$ means the number of words in document j

$|D|$ is the total number of documents in the corpus

$|\{d:t_i \in d\}|$ is number of documents in which t_i appears.

We use a parameter k to control the length of the word list. A ranked word list is generated for each document. Then we take top k words from all lists and merge these words to only one list ranked by theirs weight. This word list created by this document-centric algorithm is called *L1*. We performed some preliminary experiments, not reported here, which varied k from 1 to 110. The results reported here use $k = 30$, a value that was found to perform well.

*2) Corpus-based selection: calculates weights of words by using sum(tf)*idf*

$$W(i) = \sum_{j=1}^{n} rtf_{(i,j)} * idf_i$$

with $W(i)$ is the weight of term i;

Other parameters are calculated as same as in the first algorithm. This word list created by this corpus-centric algorithm is called *L2*.

In the second approach, i.e., *Part-of-speech* approach, we exploit the fact that words describing ontology are usually nouns. Thus, we use only words that are nouns to generate word list. These two word lists, *L1N* and *L2N* corresponding to the subset of words on lists *L1* and *L2* that are tagged as nouns using the WordNet library [19] and JWI[4] (the MIT Java WordNet Interface).

We have totally carried out different experiments for four approaches, i.e., *L1*, *L2*, *L1N*, and *L2N*. In the following sections, we will present experiment results corresponding to each approach and discuss about their performance.

IV. EXPERIMENTATION

In this section, we present experiments conducted on each component of our ontology learning framework.

[3] http://www.dmoz.org/

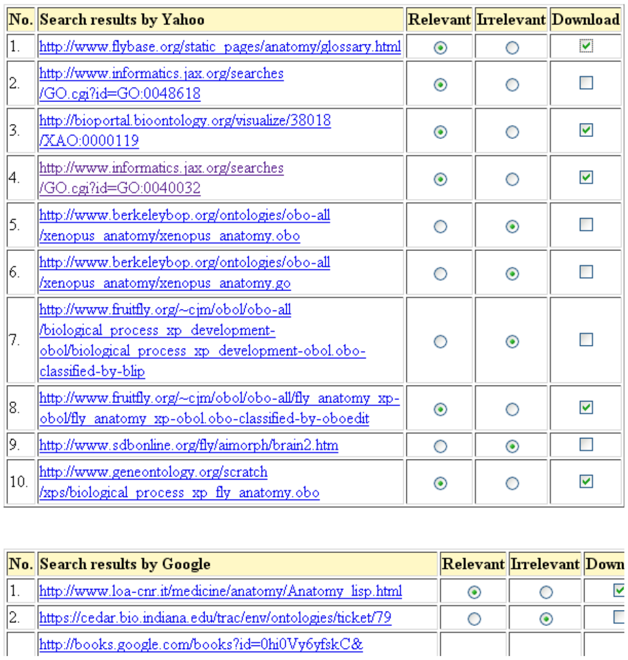
[4] http://projects.csail.mit.edu/jwi/

A. Experimentation of searching and crawling documents

The current amphibian ontology used in our experimentation is very large, containing more than 960 concepts[5]. However, due to a co-edition of this ontology among different specialists and developers in the AmphibAnat project, this current version contains many concept terms which are still not finalized (e.g., *fringe_on_postaxial_edge_of Toe_V, ventrolateral_process_of_palatoquadrate,* etc.) and noises data (e.g. *sp, aa, rr, ID_0000223*, etc.) that should be removed in the official version. Thus, the number of meaningful concepts that can be used for searching and crawling documents is decresed in our experiments. In addition, since ontology concepts are organized in hierarchy structure, there are many branches having concept names are very similar, for example the concept *foramen_acusticum_anterius* has two child concepts *foramen_acusticum_minus* and *foramen_acusticum_maius*. For this case, even we use all these concept names as keywords to look for online documents, the search results would not be better due to many duplicated words (e.g., *foramen, acusticum*) in these concepts. Therefore, we have focused on general and meaningful concept names that can be used to retrieval relevant documents in the amphibian morphology domain.

In addition, our goal is to develop techniques that can minimize manual effort by growing the ontology from a small and seed ontology, we have concentrated on experiments using a small set of keywords to search for relevant Web documents from the Internet. Thus, rather than using the ontology as input to the system, we expect to use a subset of concepts to validate our research approach. Ultimately, we hope to compare the larger ontology we build to the full ontology built by domain expert.

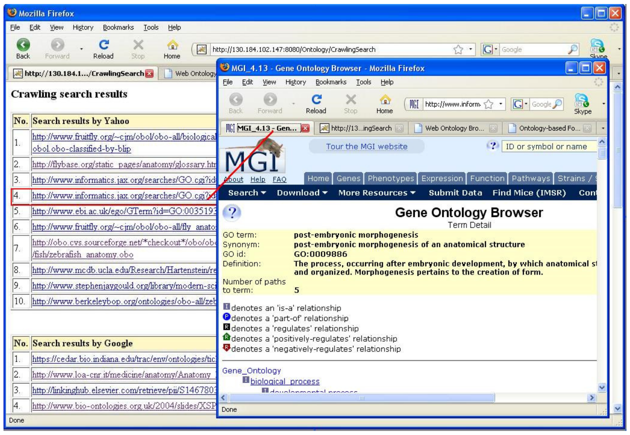
We chose a subset of 5 concepts from the amphibian ontology. From each of these concepts, we generated 3 queries with the expansion added (e.g., *"amphibian" "morphology" "pdf"*), for a total of 15 automatically generated queries. Each query was then submitted to each of the 4 search sites from which the top 10 results were requested. This resulted in a maximum of 600 documents to process. However, due to the fact that some search sites return fewer than 10 results for some queries and others are removed by our syntactic filtering and some returned documents by search engines are the same, in practice this number will be somewhat smaller. This process thus creates a very large number of hyperlinks to be analyzed, not all of which are likely to be truly relevant. Using some simple rules, these hyperlinks are automatically filtered to remove obviously irrelevant links, e.g., advertisement links and go-to-section links. The remaining links are then sent to the download module in order to retrieve the full documents. The results pages may contain documents in many formats, but we are select only HTML, pdf and text documents.



No.	Search results by Yahoo	Relevant	Irrelevant	Download
1.	http://www.flybase.org/static_pages/anatomy/glossary.html	●	○	☑
2.	http://www.informatics.jax.org/searches/GO.cgi?id=GO:0048618	●	○	☐
3.	http://bioportal.bioontology.org/visualize/38018/XAO:0000119	●	○	☑
4.	http://www.informatics.jax.org/searches/GO.cgi?id=GO:0040032	●	○	☑
5.	http://www.berkeleybop.org/ontologies/obo-all/xenopus_anatomy/xenopus_anatomy.obo	○	●	☐
6.	http://www.berkeleybop.org/ontologies/obo-all/xenopus_anatomy/xenopus_anatomy.go	○	●	☐
7.	http://www.fruitfly.org/~cjm/obol/obo-all/biological_process_xp_development-obol/biological_process_xp_development-obol.obo-classified-by-blip	○	●	☐
8.	http://www.fruitfly.org/~cjm/obol/obo-all/fly_anatomy_xp-obol/fly_anatomy_xp-obol.obo-classified-by-oboedit	●	○	☑
9.	http://www.sdbonline.org/fly/aimorph/brain2.htm	○	●	☐
10.	http://www.geneontology.org/scratch/xps/biological_process_xp_fly_anatomy.obo	●	○	☑

No.	Search results by Google	Relevant	Irrelevant	Down
1.	http://www.loa-cnr.it/medicine/anatomy/Anatomy_lisp.html	●	○	☑
2.	https://cedar.bio.indiana.edu/trac/env/ontologies/ticket/79	○	●	☐
	http://books.google.com/books?id=0hi0Vy6yfskC&			

Figure 4. Search results returned by search engines



Mozilla Firefox
File Edit View History Bookmarks Tools Help
Back Forward Reload Stop Home
http://130.184.102.147:8080/Ontology/CrawlingSearch
Google
Skype
http://130.184.1.../CrawlingSearch
Web Ontology
MGI_4.13 - Gene Ontology Browser - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Back Forward Reload Stop Home
http://www.inform
Google
Skype
MGI_4.13 - Gen...
http://13...ingSearch
Web Ontology Bro...
Ontology-based Fo...
Tour the MGI website
ID or symbol or name
MGI
About Help FAQ
Home Genes Phenotypes Expression Function Pathways Strains /
Search Download More Resources Submit Data Find Mice (IMSR) Con
Done

Crawling search results

No.	Search results by Yahoo
1.	http://www.fruitfly.org/~cjm/obol/obo-all/biological obol.obo-classified-by-blip
2.	http://flybase.org/static_pages/anatomy/glossary.htm
3.	http://www.informatics.jax.org/searches/GO.cgi?id
4.	http://www.informatics.jax.org/searches/GO.cgi?id
5.	http://www.ebi.ac.uk/ego/GTerm?id=GO:0035193
6.	http://www.fruitfly.org/~cjm/obol/obo-all/fly_anato
7.	http://obo.cvs.sourceforge.net/*checkout*/obo/obo /fish/zebrafish_anatomy.obo
8.	http://www.mcdb.ucla.edu/Research/Hartenstein/re
9.	http://www.stephenjaygould.org/library/modern-sci
10.	http://www.berkeleybop.org/ontologies/obo-all/zeb

No.	Search results by Google
1.	https://cedar.bio.indiana.edu/trac/env/ontologies/tic
2.	http://www.loa-cnr.it/medicine/anatomy/Anatomy_
3.	http://linkinghub.elsevier.com/retrieve/pii/S1467803
4.	http://www.bio-ontologies.org.uk/2004/slides/XSP

Gene Ontology Browser

Term Detail

GO term:	post-embryonic morphogenesis
Synonym:	post-embryonic morphogenesis of an anatomical structure
GO id:	GO:0009886
Definition:	The process, occurring after embryonic development, by which anatomical st and organized. Morphogenesis pertains to the creation of form.
Number of paths to term:	5

I denotes an 'is-a' relationship
P denotes a 'part-of' relationship
R denotes a 'regulates' relationship
R denotes a 'positively-regulates' relationship
R denotes a 'negatively-regulates' relationship

Gene_Ontology
I biological_process

Figure 5. Review document content before deciding to download

Figure 4 shows the returned result by each search engine. This result has been already filtered to remove irrelevant links (e.g., advertisement links and go-to-section links…) and containing only useful links that would be considered as relevant to our domain. For each returned result, we can open and see the content of this result (by clicking its URL) before deciding to download that document and classify it into the appropriate document set (c.f. Figure 5). User then can choose which documents will belong to the relevant or irrelevant set. These selected documents will be downloaded to serve the SVM classification task.

B. Experiments on Classifying and Filtering Documents

In this section, we present our experiments on training the SVM classifier to filter out the non-relevant search

[5] http://amphibanat.org/

result. The automatic nature of the corpus creation process generates a large collection of documents, not all of which are likely to be suitable for information extraction. Since extracting information from irrelevant documents would degrade the quality of the resulting ontology, it is crucial to have a filtering stage to remove irrelevant and slightly relevant documents. However, since all documents are top results retrieved from domain-relevant queries, the vocabulary overlap between the relevant and irrelevant documents is high, making this a challenging task for an automatic classifier, even one as good as SVM. Thus, the training phase is of particular importance in our work.

Using the interactive ontology-based query system described in the section III.C, we manually created a corpus of 60 relevant and 60 irrelevant Web documents retrieved by our concept-generated queries in HTML, pdf and text formats. These documents were converted into text format before using them with the SVM classifier.

1) Training the Classifier

The documents in each category, i.e., relevant and non-relevant, were divided into five subsets containing 12 documents each. For each run, two subsets are held back for testing, i.e., 12 relevant and 12 non-relevant documents, and the classifier is trained on the remaining 96 documents, 48 from each category. Thus, using five-fold cross-validation, each instance in the test collection is predicted once and the cross-validation accuracy is the percentage of documents that are correctly classified. We carry out training the classifier with and without feature selection and evaluated a variety of feature selection algorithms. For each approach, the selected features are weighted using *tf*idf* normalized by document size.

To identify important features for classification, we select those features that are most important in either the relevant set or the irrelevant set. Tokens that are appear equally frequently in both subsets are not good features for distinguishing between them. Thus, we calculated the frequency of each token in the relevant training set and also its frequency in the irrelevant training set. Finally, we calculate the *frequency difference (FD)* as the absolute difference between those two values to identify those features more strongly associated with one subset or the other. Another set of tokens that we considered as potentially important for classification is those tokens that appear only in one subset or the other. These are called the *one-subset* tokens.

We also experimented with using features that are important content descriptors for the documents, i.e., those tokens that are appear in many, but not all, documents and those which have high normalized *tf*idf* weights, meaning that they are important representations of the document contents. We call this *high distribution tokens* (HDT) selection. To run this experiment, we use parameters *m, n* and *TopN*, where *m* and *n* are the maximum and minimum number of documents containing the feature respectively, and *TopN* is the number of features selected from each document, chosen selecting the highest weighted tokens.

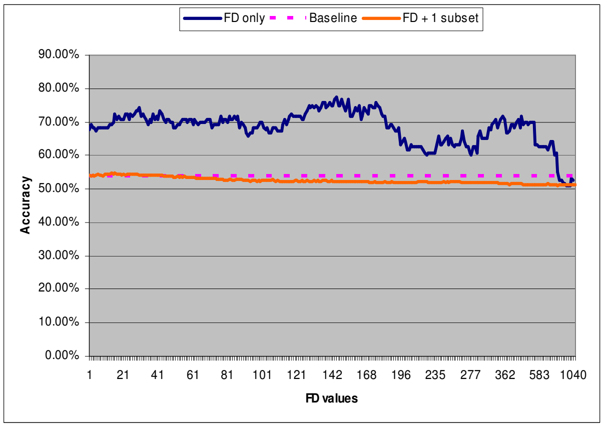


Figure 6. Average accuracy of Baseline and Feature selection with FD methods

2) Experiments

In the first experiment, we compared 3 feature selection methods:

- *No feature selection (Baseline):* We use all tokens from all documents in the training collection as features. This is our baseline against which other approaches are compared.
- *Feature selection with frequency difference (FD) only:* In this approach, we select only those tokens whose FD value is above a given threshold. We vary the FD values from 1 (all features) to 1181, at which point only 1 feature remains.
- *Feature selection with frequency difference (FD) and one-subset selection:* Features are selected as the same way and FD variation as in the above case; however we augment the feature with those tokens that appeared in only one subset.

Figure 6 shows an overall view of the baseline and the feature selection with FD methods in which we can see their accuracy with different FD values from 1 to 1181. Among these methods, the feature selection with FD only obtains high average accuracy while using just one-subset for feature selection performs worse than the baseline. Based on these experiment results, we found that feature selection with FD only performs best when using features whose frequency difference between the relevant and irrelevant sets is between 130 and 161. The peak in accuracy, 77.5%, occurred at the FD value 145, using a threshold of 0.1. The number of selected features in this case was 162.

Once we had tuned the FD method, we explored the effect of adding terms based on their frequency in the relevant set or irrelevant set. In the second experiment, we select features important representations of the document contents:

- *Feature selection with high distribution tokens (HDT):* We varied parameters values of *m, n* and *TopN* to right parameters giving the best accuracy. Experiments in this case cover all training documents distribution ranges corresponding with four values pairs *(m, n)*=(36, 12), (60, 36), (84, 60) and (96, 0), with *TopN* varies from 1 to 110.

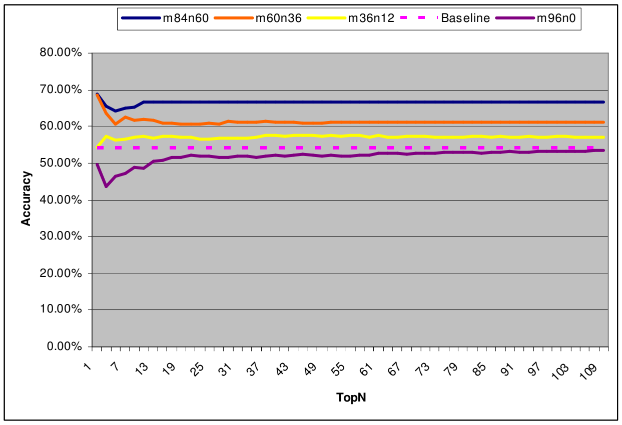


Figure 7. Average accuracy of Baseline and Feature selection with HDT methods

The second comparison (c.f. Figure 7) showed a better average accuracy result of the feature selection with high distribution tokens than the baseline. Among these lines corresponding to different parameters *m, n* and *TopN,* we found that the best result is obtained with the pair *(m, n)* = (84, 60). The results decrease if we take a range of documents having fewer features. If we choose the range covering all documents in training set and the *TopN* varies from 1 to 110, the accuracy is less than the one of the baseline as presented in the Figure 7.

C. Experimentation of Information Extraction using Text Mining

It is crucial to have a filtering stage to remove irrelevant and slightly relevant documents to the amphibian ontology. We have adopted an SVM-based classification technique trained on 60 relevant and 60 irrelevant documents collected from the Web. In earlier experiments, this spider was able to collect new documents and correctly identify those related to the domain with an average accuracy 77.5% [1].

Ultimately, the papers collected and filtered by the topic-specific spider will be automatically fed into the text mining software (with an optional human review in between). However, to evaluate the effectiveness of the text mining independently, without noise introduced by some potentially irrelevant documents, we ran our experiments using 60 documents manually judged as relevant, separated into two groups of 30, i.e., *Group_A* and *Group_B*. All these documents were preprocessed to remove HTML code, stop words and punctuation. First, we run experiment on the *Group_A* to find the case having the best result of extracting vocabulary correctly, and then we use documents in *Group_B* to validate our algorithm and compare results of these experiments.

In order to evaluate the effectiveness of the extracted words from documents, we created two *truth-lists* corresponding to the two approaches in the section 3.4. From the word list *L1* (623 words) and *L2* (623 words), after merging and removing duplicated words from these two lists, we generated the set of 507 unique words found by these two techniques. Similarly, a list of 253 unique words was generated from the lists *L1N* and *L2N*. These word lists then were judged by a human expert to classify words that are relevant or non-relevant to the amphibian morphology domain.

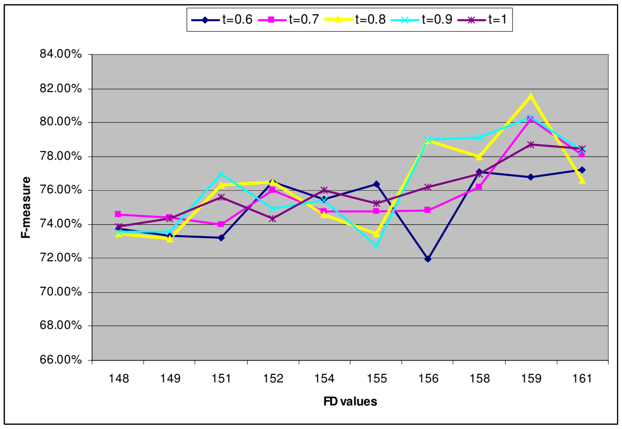


Figure 8. F-measure biased towards higher P

V. EVALUATION

We focus in this section on the performance evaluation of the two phases: SVM classification and Information extraction using text mining. For each phase, we define measures to evaluate its performance and effectiveness. We also show the comparative results and discuss the best case achieved for each phase.

A. Evaluation of SVM Classification Results

Classification effectiveness is usually measured in terms of the classic IR notions of *Precision (P), Recall (R)* and *F-measure (F)*. They can also be adapted to the case of text categorization. Denote *TP, FP, TN, FN* the number of true/false positives/negatives of returned results. These measures are calculated as following:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

$$F_{\beta} = \frac{(1+\beta^2)*P*R}{\beta^2*P+R}$$

where β allowing differential weighting of *P* and *R*.

Our experiments show that the best accuracy achieved with the FD only method is P=77.5% and R=50.7% with FD = 145. We continue to evaluate how the results achieved are varied in the best method of FD only.

Because we want to perform information extraction only on truly relevant documents, we want a metric that is biased towards high precision versus high recall. We chose to use the *F-measure* with a β value that weights precision

4 times higher than recall, i.e., *β=0.25*. We calculated the F-measure for a range around the best performing method, i.e., FD values from 130-161. For each of these FD values, we varied the SVM classification thresholds from -1 to 1 in steps of 0.1. The calculated *F-measure* results vary regularly in this range, indicating that we are getting low sensitivity with the FD method. Figure 8 shows the F-measure results for the best performing thresholds. We found that the best-performing FD approach produced an F-measure (*β=0.25)* of 81.6% with a threshold of 0.8 and FD value=159.

B. Evaluation of Information Extraction Results

In order to measure the effectiveness of our information extraction phase, we use the classic IR metrics of Precision, Recall and F-measure. We define these measures as following:

Precision (P): measures the percentage of the correct words identified by our algorithm that matched those from the candidate words.

$$P = \frac{\#\_correct\_tokens\_identified}{\#\_candidate\_tokens}$$

Recall (R): measures the percentage of the correct words identified by our algorithm that matched those from the truth list words.

$$R = \frac{\#\_correct\_tokens\_identified}{\#\_truth-list\_tokens}$$

F-measure (F): is calculated as following

$$F_{\beta} = \frac{(1+\beta^2)*P*R}{\beta^2*P+R}$$

Because we want to enhance the ontology with only truly relevant words, we want a metric that is biased towards high precision versus high recall. We chose to use the F-measure with a β value that weights precision higher than recall. From several explorations, we found that β=0.25 is an adequate value, so we used this value in our experiment.

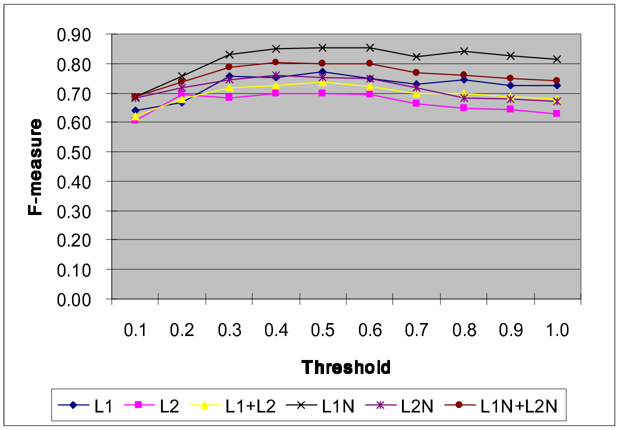


Figure 9. F-measure of the tests in Group_A

We evaluate results by comparing the candidate word lists that were extracted from the relevant documents using our algorithms with the judgments submitted by our human domain expert. We chose threshold values *t* from 0.1 to 1.0 corresponding to the percentage of top candidate words that are extracted (e.g., *t*=0.1 means that top 10% words are selected). We carried out 6 different tests corresponding to the four candidate lists, i.e., *L1*, *L2*, *L1N*, *L2N)* and two more cases *L1+L2* (average of *L1* and *L2*) and *L1N+L2N* (average of *L1N* and *L2N*) as input to our algorithm. These tests are named by their list names *L1, L2, L1+L2, L1N, L2N* and *L1N+L2N*. Figure 9 presents the F-measures achieved by these tests using various threshold values.

Figure 9 shows that the best result was achieved in the test *L1N*, using the highest weighted nouns extracted from individual documents. By analyzing results, we find that if we want a higher precision, the recall and F-measure values would decrease. We harmonize the two important values of precision and F-measures, so the best performance is achieved with a threshold *t*=0.6, i.e., the top 60% of the words (277 words total) in the candidate list are used (c.f. Table 1). This threshold produced precision of 88% and recall of 58% meaning that 167 words were added to the ontology of which 147 were correct.

Table 2 reports in more detail on the number of candidate words and how many correct words can be added to the ontology through the text mining process with the document-based selection and restricting our words to nouns only, i.e. the *L1N* test with threshold 0.6 on the validation documents, *Group_B*.

TABLE I. BEST RESULT OF THE TEST L1N (B =0.25)

Threshold	Precision	Recall	F-Measure
0.10	1.00	0.12	0.69
0.20	0.91	0.20	0.76
0.30	0.93	0.31	0.83
0.40	0.91	0.40	0.85
0.50	0.89	0.49	0.85
0.60	**0.88**	**0.58**	**0.85**
0.70	0.84	0.64	0.82
0.80	0.85	0.75	0.84
0.90	0.83	0.82	0.83
1.00	0.81	0.89	0.82

TABLE II. NUMBER OF WORDS CAN BE ADDED

Threshold	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
#candidate words	28	55	83	110	139	167	194	222	249	277
# words added	22	50	77	101	124	147	162	188	206	225

TABLE III. VALIDATED RESULT WITH GROUP_B

Threshold	Precision	Recall	F-Measure
0.6	0.77	0.58	0.70

We also observe that the top words extracted using this technique are very relevant to the domain of amphibian ontology, for example, the top 10 words are: *frog, amphibian, yolk, medline, muscle, embryo, abstract, pallium, nerve, membrane.*

To confirm our results, we validated the best performing algorithm, i.e., test case *L1N*, using the 30 previously unused relevant documents in *Group_B*. We applied the document-based selection algorithm using nouns only with a threshold value 0.6. Table 3 presents the achieved results of *P, R* and *F-measure* with threshold *t*=0.6. This shows that, although precision is a bit lower, overall the results are reproducible on a different document collection. In this case 183 words were added to the ontology of which 141 were correct

C. Discussion

Our ontology learning framework was empirically tested based on the seed amphibian ontology with retrieved Web documents by using focused crawler. An interactive system of focused crawling was created that allowed us to easily create queries from existing concepts in the ontology and submit them to Web document search engines. This system has returned many good documents since we have only taken top high-ranked search results from trusted search engines (i.e., Google, Yahoo) and through domain restricted queries. The preliminary results of relevant document classification support our hypothesis that we can use SVM to improve the identification of documents suitable for the ontology learning process. In comparison with the baseline method that used all features and produced only 53.93% accuracy, the feature selection methods generally achieve accuracy greater than 70%, with appropriate thresholds. We compared a variety of methods, and the FD method based on tokens that appear more frequently in the in either the relevant or non-relevant training sets performed the best. Adding in words that appeared in only one subset degraded performance as did a method based on the number of documents that contained the word (HDT) rather than the word frequency in each subset. When we only took tokens that occurred in many training documents, we got better accuracy than the baseline that considered all tokens from all documents, but this method's maximum accuracy was only 68.85% when tokens with the highest document counts were used. Overall, the best-performing method was FD only that achieved an accuracy of 77.5%. With a bias towards high precision, this method worked best with tokens that appeared at least 159 times more frequently in one training subset versus the other, with a high threshold of 0.8 for inclusion in the relevant class. In this case, there are 162 features used for classification which is far fewer than that total set of 40,265 features used with no feature selection. We have come up to conclude that the results are better with documents retrieved selectively by focused crawling, then filtered through the SVM classification.

For the information extraction using text mining, among four proposed approaches, we got the best results using a vector space approach with the document-based selection and restricting our words to nouns only. Overall, our algorithm produced good accuracy, over than 81% for all cases. If we restrict our candidates to only the top-weighted candidates extracted from the documents, the precision is higher but the recall decreases. In the best case, where the F-measure is maximized, the precision is 88% on the test collection. Our algorithm was also validated with another dataset (i.e. documents in *Group_B*), the precision in this case decreases to 77% which is still acceptable and does not affect significantly to the number and quality of relevant words extracted.

VI. Conclusion

In this paper, we have presented a general ontology learning framework including automated support for tasks of retrieving documents, classifying, filtering and extracting relevant information for the ontology enrichment. Our approach was empirically tested based on the seed amphibian ontology with retrieved Web documents. We have studied and implemented a focused crawler enabling us to retrieve documents in the domain of amphibian and morphology from some digital library websites or search engines. The core of our presented work is the evaluation of our SVM-based filtering technique that automatically filters out the non-relevant documents collected by the crawler so that only those most likely to be relevant are passed along for information extraction. Although the automatic collection is quite accurate, over 77.5%, this classifier could be used semi-automatically in future to allow experts to do further filtering. In the next step, only documents most likely to be relevant are passed along for information extraction.

In comparison with our previous work [1], this paper has added new content and results of the information extraction phase that enables to complete our ontology learning process. Instead of using pattern-based extraction methods, e.g., GATE tool or statistical NLP algorithms, we have applied text mining methods to identify attributes to enrich the ontology. Different experiments of text mining techniques were carried out and the precision of information extraction effectiveness which is 88% has strengthened our belief that this ontology learning process could be used semi-automatically in future to allow experts to get useful information for ontology enrichment.

VII. Future Work

Our main tasks in the future are to validate the focused crawler on a wider range of documents, experiment further with information extraction techniques to get better corpus for ontology enrichment, implement and evaluate a variety of ontology learning methods based on the domain-specific corpus.

Considering the ultimate usability of the text mining approach, it depends on the number and quality of the documents collected by the topic specific spider. In addition, although it extracts good words, these words are not matched with particular concepts within the ontology. A further pairing process, for example a matching process using WordNet vocabulary, is needed to complete the ontology enrichment process.

In future, we hope to combine this text mining approach with the one of lexical expansion using WordNet [15] to exploit the strengths of each. For example we can use WordNet pair the text mining with concepts and use the documents to identify help disambiguate the multiple senses for the concept words found in WordNet. Our other main task is to validate our approach on ontologies from other domains, to confirm that it is domain-independent. Finally, we need to incorporate the results of this work into a complete system to automatically enrich our ontology.

ACKNOWLEDGMENT

This research is partially supported by the NSF grant DBI-0445752: Semi-Automated Construction of an Ontology for Amphibian Morphology.

REFERENCES

[1] H. Luong, S. Gauch and Q. Wang, "Ontology-based Focused Crawling", International Conference on Information, Process, and Knowledge Management (eKNOW 2009), Cancun, Mexico, Feb. 1-7, 2009, pp. 123-128.

[2] E. Agirre, O. Ausa, E. Havy and D. Martinez, “Enriching Very Large Ontologies Using the WWW”, *ECAI 1st Ontology Learning Workshop*, Berlin, August 2000.

[3] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web”, *In Scientific American,* 2001, pp. 35-43.

[4] P. Buitelaar, P. Cimiano and B. Magnini, “Ontology Learning from Text: Methods, Evaluation and Applications”, *IOS Press (Frontiers in AI and applications*, vol. 123), 2005.

[5] S. Chakrabarti, M. Berg and B. Dom, “Focused crawling: a new approach to topic-specific Web resource discovery”, *Computer Networks: The International Journal of Computer and Telecommunications Networking*, Elsevier North-Holland, May 1999, **31**(11-16), pp. 1623 – 1640.

[6] C-C. Chang and C-J. Lin, “LIBSVM : a library for support vector machines”, 2001, Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[7] H. Cunningham, D. Maynard, K. Bontcheva, and V. Tablan. “GATE: A framework and graphical development environment for robust NLP tools and applications”, *Proceedings of Computational Linguistics*, 2002.

[8] M. Ester, M. Gross and H.-P. Kriegel, “Focused Web Crawling: A Generic Framework for Specifying the User Interest and for Adaptive Crawling Strategies”, *27th Int. Conf. on Very Large Databases*, Roma, Italy, 2001.

[9] S. Gauch, J. M. Madrid, S. Induri, D. Ravindran, and S. Chadlavada, “KeyConcept: A Conceptual Search Engine”, Center, Technical Report: ITTC-FY2004-TR-8646-37, University of Kansas.

[10] A. Gómez-Pérez, and D. Manzano-Macho, “A survey of ontology learning methods and techniques”. Deliverable 1.5, IST Project IST-2000-29243 - OntoWeb, 2003.

[11] T. Gruber, “Towards principles for the design of ontologies used for knowledge sharing”. Int. J. of Human and Computer Studies, 1994, (43), pp.907–928.

[12] A.H. Tan, “Text mining: The state of the art and the challenges”. In Proceedings of the Pacific Asia Conf on Knowledge Discovery and Data Mining PAKDD'99 workshop on Knowledge Discovery from Advanced Databases, pages 65-70, 1999.

[13] A. Hotho, A. Nürnberger, and G. Paaß, “A Brief Survey of Text Mining”. LDV-Forum, 20(1):19–62, .2005.

[14] T. Joachims, “Text categorization with support vector machines: learning with many relevant features”, *Proceedings of the 10th ECML-1998,* pp. 137–142.

[15] H.Luong, S.Gauch and M.Speretta, “Enriching Concept Descriptions in an Amphibian Ontology with Vocabulary Extracted from WordNet”. The 22nd IEEE Symposium on Computer-Based Medical Systems (CBMS 2009), New Mexico, USA, August 2-4, 2009, pp 1-6.

[16] A.M. Maglia, J.L. Leopold, L.A. Pugener and S. Gauch, “An Anatomical Ontology For Amphibians”, *Pacific Symposium on Biocomputing*, 2007, (12), pp.367-378.

[17] A. Maedche and S. Staab, “Ontology Learning for the Semantic Web”. *IEEE Intelligent Systems, Special Issue on the Semantic Web*, March 2001, **16**(2), pp. 72 - 79.

[18] A. Maedche, G. Neumann and S. Staab, “Bootstrapping an Ontology-Based Information Extraction System”. Studies in Fuzziness and Soft Computing, Intelligent exploration of the web, Springer, 2003, pp.345 – 359.

[19] G-A. Miller, “WordNet: a lexical database for english”, Comm. ACM, Vol 38, No. 11, pp. 39-41, 1995.

[20] B. Omelayenko, “Learning of ontologies for the Web: the analysis of existent approaches”, *Proceedings of the international workshop on Web dynamics*, London, 2001.

[21] M. Shamsfard and A.A. Barforoush, “The State of the Art in Ontology Learning”, *The Knowledge Engineering Review*, Cambridge Univ. Press, 2003, 18(4), pp. 293 – 316.

[22] M. Speretta and S. Gauch, "Using Text Mining to Enrich the Vocabulary of Domain ontologies", 2008 IEEE/WIC/ACM Int. Conference on Web Intelligence, Sydney, Australia, Dec. 9-12, 2008, pp. 549-552.

[23] S. T. Dumais, J. Platt, D. Heckerman, and M. Sahami, “Inductive learning algorithms and representations for text categorization”, *Proceedings of CIKM-98,* pp. 148–155.

[24] V. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag, 1995.

[25] M.W. Berry, “Survey of Text Mining”, Springer-Verlag New York, Inc., Secaucus, NJ, 2003.

[26] Y.Yang and J.O. Pederson, “A comparative study on feature selection in text categorization”, (*ICML)*, 1997.

[27] Y. Yang, J. Zhang and B. Kisiel, “A scalability analysis of classifiers in text categorization”, *Proceedings of 26th ACM SIGIR Conference*, July-August 2003, pp 96-103.

www.iariajournals.org

International Journal On Advances in Intelligent Systems

ICAS, ACHI, ICCGI, UBICOMM, ADVCOMP, CENTRIC, GEOProcessing, SEMAPRO, BIOSYSCOM, BIOINFO, BIOTECHNO, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS

issn: 1942-2679

International Journal On Advances in Internet Technology

ICDS, ICIW, CTRQ, UBICOMM, ICSNC, AFIN, INTERNET, AP2PS, EMERGING

issn: 1942-2652

International Journal On Advances in Life Sciences

eTELEMED, eKNOW, eL&mL, BIODIV, BIOENVIRONMENT, BIOGREEN, BIOSYSCOM, BIOINFO, BIOTECHNO

issn: 1942-2660

International Journal On Advances in Networks and Services

ICN, ICNS, ICIW, ICWMC, SENSORCOMM, MESH, CENTRIC, MMEDIA, SERVICE COMPUTATION

issn: 1942-2644

International Journal On Advances in Security

ICQNM, SECURWARE, MESH, DEPEND, INTERNET, CYBERLAWS

issn: 1942-2636

International Journal On Advances in Software

ICSEA, ICCGI, ADVCOMP, GEOProcessing, DBKDA, INTENSIVE, VALID, SIMUL, FUTURE COMPUTING, SERVICE COMPUTATION, COGNITIVE, ADAPTIVE, CONTENT, PATTERNS

issn: 1942-2628

International Journal On Advances in Systems and Measurements

ICQNM, ICONS, ICIMP, SENSORCOMM, CENICS, VALID, SIMUL

issn: 1942-261x

International Journal On Advances in Telecommunications

AICT, ICDT, ICWMC, ICSNC, CTRQ, SPACOMM, MMEDIA

issn: 1942-2601