Ontology for a R+D+I Centre to organize, retrieve and share information and knowledge resources for personal as well as corporate use: Suricata platform.

José C. Nelson, Manuel Galán, Antonio Ocón, Enrique Rubio

CICEI (Center for innovation in the information Society)-University of Las Palmas de Gran Canaria- Spain

jnelson@cicei.ulpgc.es, mgalan@cicei.com, ocon@cicei.com, rubio@cicei.com

Abstract

Employees in a R+D+i Centre face the challenge of retrieving, indexing, and transforming information from large and heterogeneous databases with a lack of organization.
Tools, like a Corporate and Personal Management System, ought to be devised to help employees to manage information and knowledge resources more efficiently and more effectively.

We present the lay out of the conceptual framework for the information architecture of the Suricata platform. In this platform we will not consider the content of the documents rather we build a layer of meta-information.
The technologies that we use are agent-based systems and ontologies as a controlled language.

Keywords Ontology

1. Introduction

In recent years there has been an unprecedented surge in global information sources of an astronomical magnitude.
Employees in a R+D+i Centre face the challenge of retrieving, indexing, and transforming the ever-increasing and ever-changing information, perhaps of an astronomical magnitude. An efficient information retrieval system can be undermined by database size, heterogeneity (information available may be formed by documents without a common structure, most of them expressed in natural language and without semantics as the corporate memory could be) and lack of organization (information scattered without any criteria).
The R+D+i centres have knowledge that is really large and complex, therefore it becomes fundamental to exploit it, providing knowledge workers with effective means for communicating their knowledge for the benefit of the whole R+D+i centre. Since there are a lack of time and resources in these sorts of environments, tools ought to be devised to help employees to manage information and knowledge resources more efficiently and more effectively. Furthermore, personal and corporate productivity would be improved and allowing for greater depth and focus in their respective duties.

2. Suricata platform

As a challenge to the new socioeconomic context, the work based on knowledge, we have developed a socio-technical model of innovation to be applied within organizations to do the transition to e-knowledge economy, called Suricata. Based on this model we are developing Suricata platform whose architecture is formed by several functional layers [5].

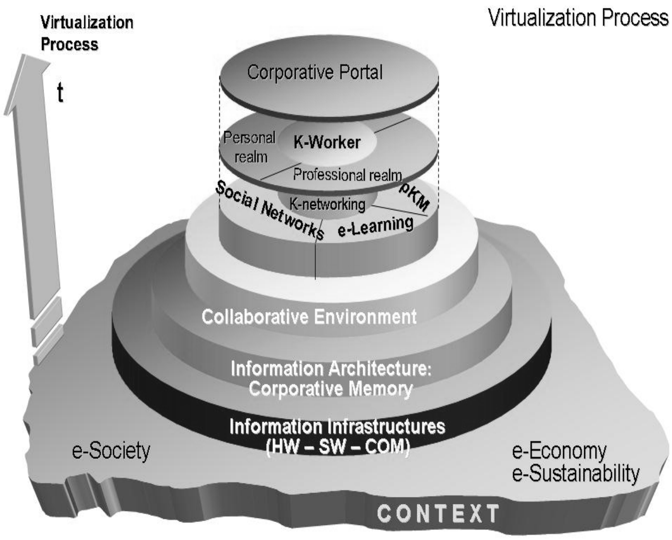

Fig. 1: Suricata Model

In order to be implemented, we started from a cooperative environment called IDESKTOP, developed in our R+D+i Centre CICEI at the University of Las Palmas de Gran Canaria.
That cooperative environment will be the foundation and point of departure for our final goal: a Corporate and Personal Management System.

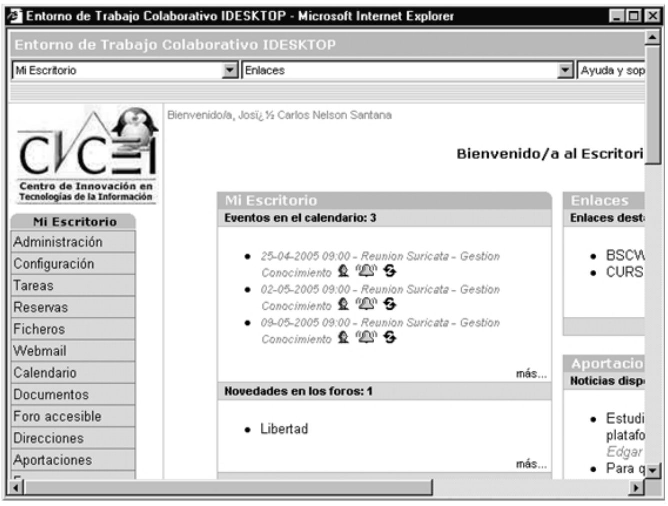

Fig.2: Idesktop

The reason this goal has been chosen is that the existing approaches do not usually take into account individual aspects and only target the organization’s needs. There are some solutions to help an employee to manage information and knowledge resources.

The most common solution is an archiving hierarchy as well as a content management system.
The design of an archiving hierarchy is based on combining elements of organizational and thematic categories. A drawback is that few users have the skills to produce and manage effectively the existing system that will allow them to search and retrieve information resources.

On the other hand the content management systems are useful to handle a large amount of documents which are shared by the employees in a R+D+i Centre. The documents have to be filed, protected, etc. in order to avoid problems such as multiple copies of the same document. The content management systems are not useful to give insightful ways of linking the documents and the knowledge associated with them.

Here we give a brief overview of the lower layers of this platform which has been proposed before [10].
The Information Infrastructure is a basic layer and it contains the infrastructure formed by hardware, software and communications which are needed to work. The main features of this layer are an architecture that is flexible and scalable, which contains a fault-tolerant system and Open Source philosophy.
The next layer contains the Information Architecture and the Corporate Memory. The last one is the knowledge base of the organization and it backs up its knowledge. The management of the Corporate Memory is a key component in organizations like R+D+i Centres because any user should be able to work with the knowledge base. Consequently, there are interactions that solve problems and perform tasks which require the use of a combination of knowledge techniques and also a management orientated towards processes.
The organizations devise strategies focused to fulfill objectives which have been set in advance and these strategies are reached through the use of process
We consider several types of tasks in a process: the *trade processes* which are intense with respect to the knowledge and generate the outputs of production, the *processes of knowledge* which support the knowledge interchange between departments and the trade process and the *processes related to knowledge management*, responsible for the control and support of the knowledge base.

3. Information architecture

In this paper, we present a conceptual framework for the layer of information architecture of Suricata model and an attempt at an initial lay out design.

Some of the architectural features proposed are the following:
-Emergence: emergence occurs when the activities of the parts do not simply sum to give activity of the whole or the whole is more than the sum of its parts [6]

- Non-linear: the number of possible user´s interaction is a non-linear function which depends on a finite number of e-contents, processes and users

$$I = C\,P\,U \qquad (1)$$

where

C: e-contents
P: processes
U: users

If we consider these features, it would be impossible to make a prediction about the whole set of a user’s interactions when he interfaces with the corporate and personal management System.

- Multi-layer structure: several layers of content which will depend on personal, group or centre level context

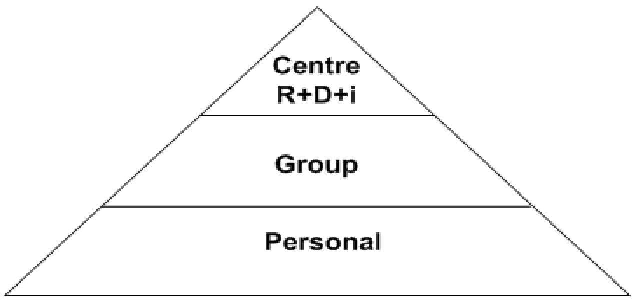

Fig.3: Multi-layer structure

- Network flows: information flows among employees.

- Categorization: the basic cognitive process of arranging into categories. The human brain processes huge quantities of information and it is able to cope with so many inputs owing to fact that it categorizes them all. The brain looks for what is new, what is different and what has changed. Then the new information is matched to the categories that exist. Finally, the new input is dropped into the category that is most similar. Information systems also need organization if users want to make sense of their contents. The categorization attempts to solve the problems indicated before such as information gathering, organizing and disseminating.

Since this multilayer structure is a complex adaptive system we have Suricata as an adaptive information architecture that greatly improves navigation through this system thus helping knowledge workers to share their ideas and knowledge about their work with other employees in order to produce value.
Agent-based systems are software that provides a flexible means of building an adaptive infrastructure to perform specific tasks on behalf of their users [13]. They can improve the precision (the ability of the system to retrieve relevant information) and recall (the ability not to retrieve non-relevant information) of knowledge on the R+D+ i Centre.

If we use the Semantic Web as model which “includes documents, or portions of documents, describing explicit relationships between things and containing semantic information intended for automated processing by our machines”, we could use intelligent agents [5] to lead structured knowledge extraction instead of simple information retrieval.
The agents are implemented in Java language using a platform of multi-agent systems such as JADE (Java Agent DEvelopment Framework) which is a software development framework aimed at developing multi-agent systems and applications that complies with the FIPA specifications for intelligent agents. It includes two main products: a FIPA-compliant agent platform and a package to develop Java agents. One package allows using a graphical user interface that is show in fig.3.

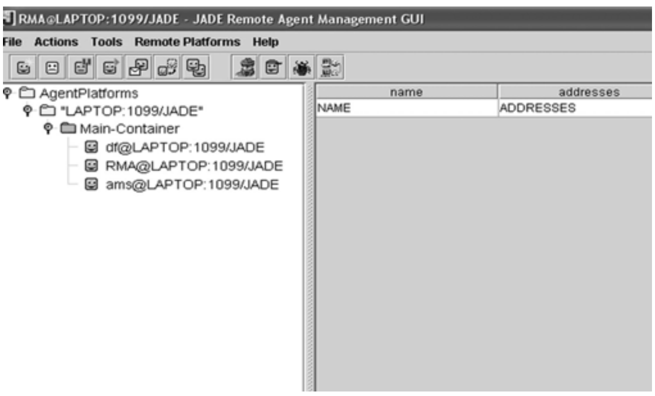

Fig.4: Graphical user interface

In traditional information retrieval models the content of the documents are treated by means of techniques based on generalized vector, fuzzy set theory or Bayesian networks.
In Suricata platform we should not consider the content of the documents rather we should build a layer of meta-information.

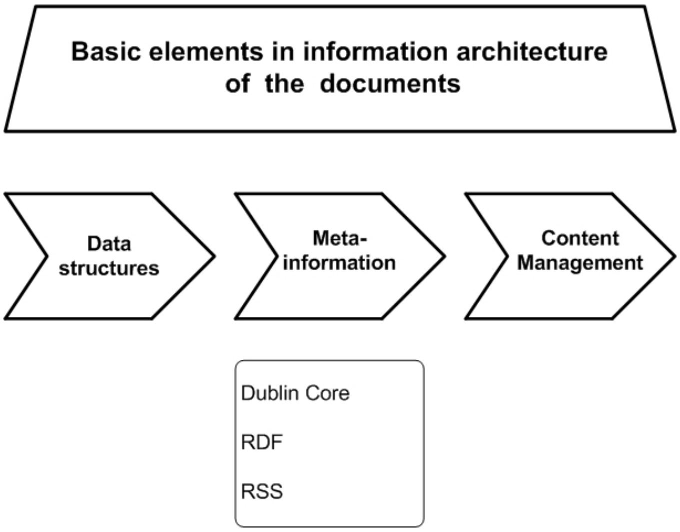

Fig.5: Basic elements in information architecture of the documents

Therefore meta-tags ought to be used in documents produced by the organization in order to form this set of meta-information. As we know they are used by the agents to index the content of the pages which is essential to search and retrieve the information. The meta-tags are under the specification of W3C and Dublin Core.
We use **Dublin Core** metadata according ISO standard 15836-2003, RSS and RDF according to W3C recommendation.
The Dublin Core Metadata Initiative is an open forum engaged in the development of interoperable online metadata standards that support a broad range of purposes and business models. DCMI's activities include consensus-driven working groups, global conferences and workshops, standards liaison, and educational efforts to promote widespread acceptance of metadata standards and practices.

The following Dublin Core metadata can be used in each document stored in the Corporate and Personal Management System in addition to user’s devised meta-tags.

Label:	Title
Definition:	A name given to the resource.
Label:	Description
Definition:	An account of the content of the resource.
Label:	Creator
Definition:	An entity primarily responsible for making the content of the resource.
Label:	Subject and Keywords
Definition:	A topic of the content of the resource.

Label:	Publisher
Definition:	An entity responsible for making the resource available
Label:	Contributor
Definition:	An entity responsible for making contributions to the content of the resource.
Label:	Date
Definition:	A date of an event in the lifecycle of the resource.
Label:	Resource Type
Definition:	The nature or genre of the content of the resource.
Label:	Format
Definition:	The physical or digital manifestation of the resource.
Label:	Resource Identifier
Definition:	An unambiguous reference to the resource within a given context.
Label:	Source
Definition:	A Reference to a resource from which the present resource is derived.
Label:	Language
Definition:	A language of the intellectual content of the resource.
Label:	Relation
Definition:	A reference to a related resource.
Label:	Coverage
Definition:	The extent or scope of the content of the resource.
Label:	Rights Management
Definition:	Information about rights held in and over the resource.

RSS is a file format based on XML(Extensible Markup Language) and RDF (Resource Description Framework) is essentially a data-model that permits the creation of publication channels that are easy to read with special programs that do not need to be linked to page.

The other core feature is to have a language controlled by ontology to be used in the development of the information system to recover data from heterogeneous databases as the corporate memory and documents produced in business processes of our Centre.
In last years, the world ontology has been used in the Knowledge Engineering community. Some definitions have given by different authors to define ontology. Following Gruber [2], we use the term to indicate that *"Ontology is an explicit specification of a conceptualization "*

But in our proposal we use the followed definition [9]

> *The ontology defines the basic terms and relations including the vocabulary of a topic area as well as the rules to combine terms and relations to define extensions to the vocabulary.*

The ontology is used for inferring knowledge from the vocabulary. The ontology makes it possible to catalogue the contents of corporate memory, the documents generated and also to assign semantic attributes in order to represent the various elements of information.

The ontologies have to be devised [1] and we need objective criteria to lead and test the designs. A set of design criteria for ontologies has been proposed [3] and we have used some of them:

1. Clarity: An ontology should effectively communicate the intended meaning of defined terms.
2. Coherence: It should confirm inferences that are consistent with the definitions.
3. Extendibility: An ontology should be devised to anticipate the uses of the shared vocabulary. So that, it should be able to make up new terms based on the existing vocabulary.

Ontology design does not have an immediate solution and as other design problems will call for tradeoffs among the criteria.
There are some types of ontologies and some ways of categorizing them. In the design, the ontologies for a R+D+i Centre we have used the classification of ontologies according to their level of dependence on a particular task or point of view [4]

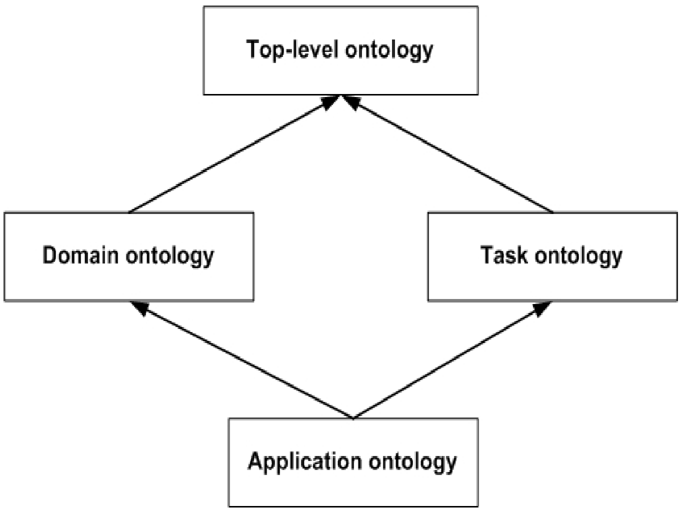

Fig.6: Classification of ontologies

Here we define the ontologies more relevant in our R+D+i Centre.
Top-level ontologies describe very general concepts and provide general notions under which all root terms in existing ontologies should be linked.

Task ontologies describe the vocabulary related to a generic task or activity by specializing the terms in the top-level ontologies.

Domain ontologies lay on vocabularies about concepts within a domain and their relationships, about the activities taking place in that domain, and about theories and elementary principles governing that domain.

Application ontologies contain all the definitions needed to model the knowledge required for a particular application. They usually spread and make more specific the vocabulary of the domain and task ontologies for a given application.

We have created a specific domain ontology in order to provide vocabularies about concepts within the domain of the R+D+i Centre and their relationships to the activities taking place in this domain.

Furthermore, the ontologies would provide a model to stand for the organisation. We indicated before that few users can manage an archiving hierarchy properly. As a consequence, we also use the ontologies as a document organisation tool.

4. Ontology developed

We present as an example a process very common in our R+D+i Centre. Most of their employees teach at Technical Faculties as Engineering or Computing Science and they direct the Final Projects. This is the process in which most of the people is involved but has not been standardized. Because of this, we developed the ontology of the "Final Project Management".
An analysis of the processes of each individual will be completed using conceptual maps and an approximation of the CommonKads methodology [7].

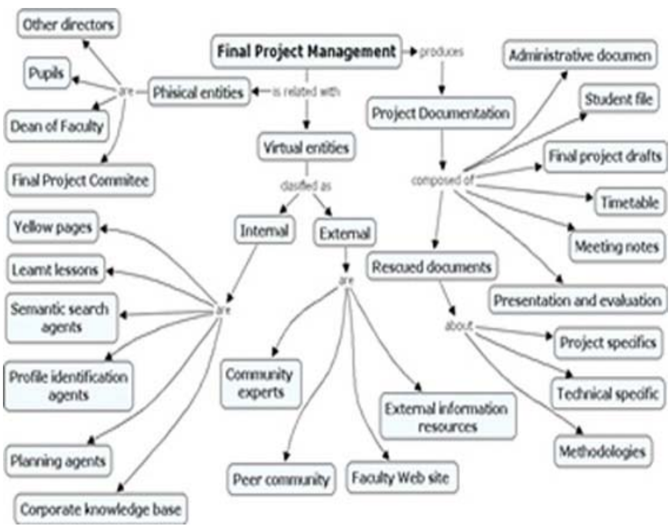

Fig. 6: Conceptual map of the process "Final Project Management"

As we see the use of tools such as conceptual maps can help to design the ontology more quickly. We have used the Protege ontology editor from Stanford University which is congruent with the Open Source philosophy.

The superclasses are People, Virtual, Documentation and Organization. The overall numbers of classes are fifteen: Pupil, Director, Dean of Faculty, Semantic Search Agent, Profile Identification Agent, Corporate Knowledge Base, Peer Community, Community Experts, External Information Resources, Yellow Pages, Faculty Website, Learnt Lessons, Planning Agents, Meeting Notes, Timetable, Meeting Notes, Administrative Documents, Used Documents, Student File, Final Project, Presentation and Evaluation, and Final Project Committee.
It is necessary to define the slot of each class. The next step is entering instances. These are the actual data in the knowledge base.
In Fig.7, there is a partial view of Protege Interface and we can also see the slots of class Pupil.

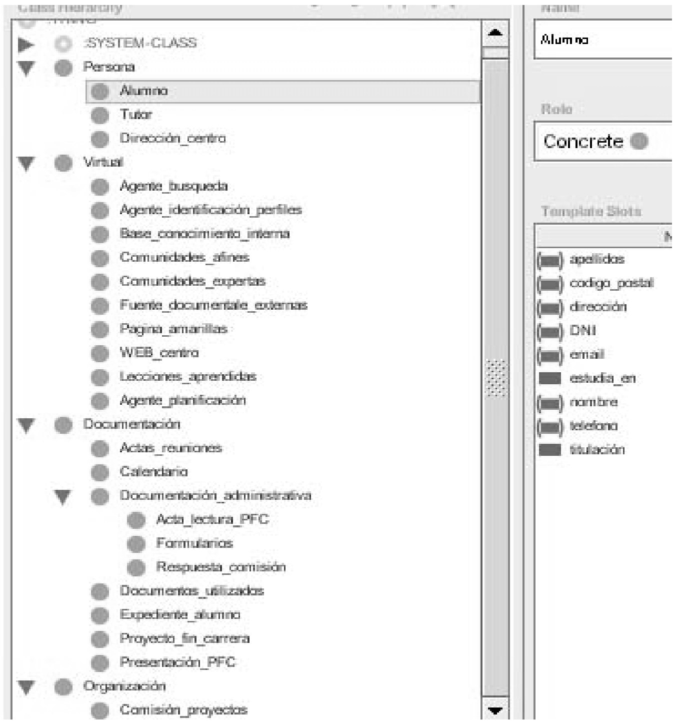

Fig.7: View of the Final Project Management ontology

The last step is creating the queries in order to locate all instances that match criteria we specify.

In order to check the scope of the ontology we have to lay out a list of questions that the knowledge base on the ontology should be able to answers. They are called competency questions []

5. Conclusion

In this paper, we have described a conceptual framework for the layer of information architecture of Suricata model. This proposal is based in three main features, the first is using intelligent agents, the second is building a layer of meta-information by using meta-tags and the third is using a language controlled by an ontology which is also used to categorize the information. In this paper, we have described an ontology-centre approach as a tool to organize, retrieve and share information and knowledge resources in a R+D+i Centre. We have tried to highlight the key issues in designing an ontology for one of the processes developed in our Centre.

Acknowledgements

Project funded by Ministry of Education and Science of Spain (National Plan R+D+i) according to the Research Project reference TSI 2004-05949 financed jointly with EU Feder funds.

References

[1] Gómez-Pérez, A., Fernández-López, M. and Corcho, O. " Ontological Engineering", *Springer-Verlag*, (2004).

[2] Gruber, T.R. "A Translation Approach to portable ontology specification", *Knowledge Systems Laboratory, Stanford University, Technical Report KSL 92-7*, (1993).

[3] Gruber, T.R. "Toward principles for the design of ontologies used for knowledge sharing", *Knowledge Systems Laboratory, Stanford University, Technical Report KSL 93-04,* (1993).

[4] Guarino, N. "Formal Ontology in Information Systems", *IOS Press*, (1998).

[5] Hendler, J. "Agents and the Semantic Web", *IEEE Intelligent Systems*. March-April (2001).

[6] Holland, J. "Emergence from chaos to order*", Oxford University Press,* (2000).

[7] Marrero, S.,Nelson, J.C.,Galen, M., Ocon, A. and Rubio, E." Metodología para organizar, recuperar y compartir recursos de información y conocimiento en un Centro I+D+i en la plataforma Suricata". Paper presented at *IADIS* 2005

[8] Mizoguchi, R. and Ikeda, M. "Towards Ontology Engineering". *Technical Report AI-TR-96-1, I.S.I.R., Osaka University.*

[9] Neches, R.., Fikes, R.E., Fininn, T., Gruber T.R., Senator T.,and Swartour, W.R., "Enabling technology for knowledge sharing". *AI Magazine 12(3):36-56*,(1991).

[10] Rubio, E., Ocón, A.,Galán, M., Marrero, S. and Nelson, J.C."A personal an corporative process-oriented knowledge manager: Suricata model". *Eunis* , (2004).

[11] Syed, J. "An adaptive framework for knowledge work". *Journal of Knowledge Management*, Vol. 2. Number 2. December (1998).

[12] Uschold, M. and Gruninger, M. "Ontologies: principles, methods and applications". *Knowledge Engineering Review*. Volumen 11 Number 2, June 1996.

[13] Wooldridge, M. "An introduction to MultiAgent Systems". *Wiley* (2001).