ONTOLOGY-BASED LAND DEGRADATION ASSESSMENT FROM SATELLITE IMAGES

E. Tomai [a, *], I. Herlin [b], J.-P. Berroir [b], P. Prastacos [c]

[a] NTUA, School of Rural and Surveying, 15780 Zografos Campus Athens, Greece – etomai@mail.ntua.gr
[b] INRIA, BP105, 78153 Le Chesnay Cedex, France – (Isabelle.Herlin, Jean-Paul.Berroir)@inria.fr
[c] FORTH, Institute of Applied and Computational Mathematics, 711 10 Heraklion, Greece – poulicos@iacm.forth.gr

Commission II, WG II/6

KEY WORDS: Land Use, Change Detection, Understanding, Updating, Knowledge Base, Parameters

ABSTRACT:

In this paper, we introduce the idea of documenting operational chains for land degradation assessment using ontologies. We believe that this will help end-users in better understanding the land degradation characteristics and evaluate the results of the assessment process. Since the application domain is wide, various operational chains for land degradation assessment and their associated documentation exist, according to different options. This parameterization causes the development of different ontologies, which, nonetheless are to a certain extent linked because of the common software components of the corresponding operational chains. We therefore propose a hierarchical structure of these ontologies; so that several requirements such as understanding of expert knowledge interconnections and application domain variety, documentation, assimilation of new expert knowledge, and reusability of software components become feasible.

1. INTRODUCTION

The objective of the research is the use of ontologies in land degradation assessment applications monitored by satellite images. The motivation behind the use of ontologies, in such an endeavour, is to help dissemination and usage of software components among a range of scientists and users of environmental applications. We believe that the ontology-based documentation of the operational chains end-users have to pursue will facilitate both the assessment of land degradation and the understanding of its characteristics. Actually, ontologies enable the community of users of an operational chain to understand how the various processes of the chain are interrelated since they document the various parameterisations needed in this type of applications. The main quality of ontologies is that they make problem-solving knowledge explicit to end-users, by providing a common vocabulary and by giving to it a clear-cut meaning.

Among the different types of ontologies that are presented in Guarino (1998), and Gomez-Perez and Benjamins (1999), we isolate herein: the task, method and application ontologies. A task-ontology provides a systematic vocabulary of the terms used to solve problems associated with tasks. A method-ontology provides definitions of the relevant concepts and relations used to specify a reasoning process in order to achieve a particular task And finally an application-ontology contains the necessary knowledge for modelling a particular application. As it can be induced from these definitions, a method ontology can be considered as a subcategory of a task ontology.

Since the application domain of land degradation is wide, various operational chains for land degradation assessment and their associated documentation exist, according to different options. These options comprise of: different data sources (AVHRR, MODIS,…) with various spectral, spatial and temporal characteristics; different applicative goals: desertification, deforestation, erosion, etc.; different sites of interest, such as Brazil, South Africa etc.; and different algorithms with associated data structures, e.g. representing a land use label by membership functions, or as a collection of spectral samples, etc.

For each previously mentioned option an operational chain is defined according to a specific parameterisation of the software components. A comprehensive documentation is required for each of these operational chains, since they are most often developed by scientists, but operated by end-users. The latter need to process them easily and as automatically as possible, while at the same time be able to have control and complete understanding of the intermediate processes so that they can interpret the results. These operational chains are, to a certain point, linked because they have several software components in common.

As mentioned, each operational chain is defined according to a parameterization that reflects a specific set of options. This parameterization process underlies the development of different ontologies, which, nonetheless, can be organized in a specific configuration, corresponding to an oriented and hierarchical tree structure. The ontologies of the operational chains constitute the lower level of this tree. Within this configuration, ontologies at one level are directly obtained from those at the previous level after defining some of the software components and parameterizations. Therefore, this hierarchical structure enables automatic derivation of ontologies compatible with the chosen software components and parameterizations (Oberle, 2005) that constitute an operational chain, for each type of application under a given context.

Additionally, through the proposed structure we answer, as it will be further shown in the paper, the questions of: a) how new knowledge, when and if acquired, can be incorporated in an existing operational chain and therefore update it, and b) how a more detailed operational chain can be developed using this new knowledge.

Finally, we seek, herein, to prove that the organization of ontologies in a hierarchical structure meets several other requirements. More specifically, it: a) provides an understanding of the underlying variety of the available operational chains, b) establishes interconnections so that different parts of knowledge, belonging to the same context, become coherent and shared by scientists and end-users, and c) enables end-users to select, among a variety of choices, the operational chain that meets best their needs.

2. DEFINING AND EXPLAINING THE TREE

In this section, we define what is included in the tree (figure 1). At the top (level 0), there is the ontology which corresponds to the learning phase. This ontology represents problem-solving knowledge and the available methodologies for executing the required tasks. At lower levels are the ontologies for specific environmental applications (e.g., deforestation or desertification) parameterized for different datasets (e.g., MODIS or NOAA) (corresponding to levels 1 and 2). At the lowest level are fully parameterized ontologies in the context of a specific geographic area (each corresponding to a specific operational chain; for instance an ontology for identifying deforestation areas by analyzing MODIS data for the Taquari basin in Brazil as the one presented in section 4). Thus, end-users, using the appropriate ontology are able to retrieve the operational chain that fits the specific application context, data, location, etc.

How does the tree structure helps meeting the objectives presented in the introduction? In what follows we discuss the main benefits of such a structure. First, when having such a hierarchical structure scientists are able to understand the interdependencies among different parts of knowledge, by going from the upper to the lowest level. This is so, because the processing chains (and, therefore, the corresponding ontologies), due to the software components that underlie them, become more and more specialized, parameterized and calibrated to account for the specific data, location context etc. Hence, scientists are able to reason about the different parts of knowledge contained at each level of the tree structure. Within this type of structure, the ontologies, but more importantly the different underlying operational chains are coherent. Since the relations that exist in the tree structure are well known and articulated, it becomes feasible to analyze the results of a land degradation assessment process according to the algorithmic choices represented by the path from the root to the lowest level of the tree; where the operational chains are located.

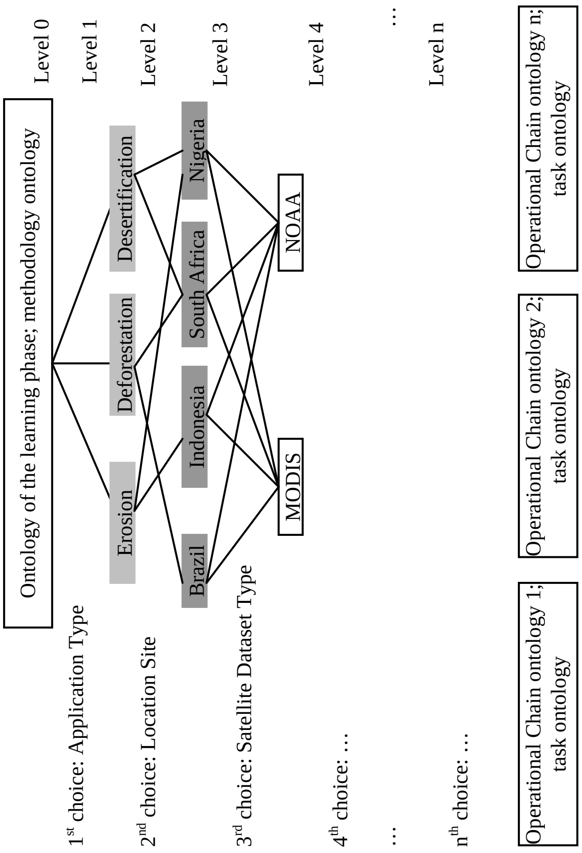
Level 0
Level 1
Level 2
Level 3
Level 4
…
Level n
Ontology of the learning phase; methodology ontology
Desertification
Deforestation
Erosion
Nigeria
South Africa
Indonesia
Brazil
NOAA
MODIS
Operational Chain ontology n; task ontology
Operational Chain ontology 2; task ontology
Operational Chain ontology 1; task ontology
1st choice: Application Type
2nd choice: Location Site
3rd choice: Satellite Dataset Type
4th choice: …
…
nth choice: …

Figure 1: The tree structure of ontologies.

A second advantage is the ability to add new parts of knowledge. If, at one point, scientists perfect their knowledge about the application domain by refining choices among methodologies or types of data, then it is just necessary to add new parts in the structure (most probably a complete intermediate level) to represent the newly acquired components of knowledge. Each leaf of the tree corresponds to a set of choices, which has to be made in the process of defining operational chains for end-users. Therefore, the refinement of the choices leads to inserting additional leaves or even levels in the tree.

A third advantage is that the tree structure enables updating of the operational chains. Suppose end-users employ an operational chain (therefore, the associated ontology) and at one point they decide to make modifications to the chain, due to their experience. In this case, ontologies have to be modified accordingly. Then these modifications should be transmitted on the tree structure; upward, and then downward to be shared by others. Consequently the key-idea, here, is to update the ontologies both horizontally and vertically along the tree structure.

One consequence of the previous points is that ontologies included in the structure and their associated operational chains can be reusable. This practically means that if an operational chain is defined for one application on one geographic location, it will be easy to adapt it for another geographic location and/or for another application and to derive the corresponding ontology, because, in such case, problem-solving knowledge is significantly the same. Hence, software components are also reused, allowing reducing programming effort from the developer’s point of view.

3. ISSUES IN DEVELOPING THE HIGHEST LEVEL ONTOLOGY

This section explains the development of the ontology at the highest level and the theoretical issues met during its development. This highest level concerns the general undertaken methodology to be followed during the learning phase, that is, a data analysis phase whose results serve as inputs to operational land degradation monitoring and assessment. The corresponding ontology is generic and includes all the necessary components of a method-ontology (Chandrasekaran et al., 1998 and Rajpathak & Motta, 2004) such as: Initial state, Goal(s), Tasks, Result(s), Methodology, Parameters, Constraints, Input(s), Output(s).

In developing ontologies for describing a methodology or a sequence of tasks, we face several theoretical issues of the learning phase, which are presented in the following subsections. The initial choices are fully up to the users, supported or not by different objective criteria, which allow justifying and quantifying the preferred alternative (e.g. choice of geographic site, of classification method). Further choices are supported by automated tasks, as the determination of the low-resolution nomenclature. The latter choices will constitute the core of the learning phase, the purpose of which being to end-up with the full parameterization of an operational chain.

3.1 The initial parameterization

The initial parameterization consists of setting the application type and its geographic location. Therefore:

1. The first choice concerns the application type. The application refers to different land degradation assessment cases such as deforestation, erosion, or desertification. This choice is simple. The user knows in advance the type of the application.
2. The second choice concerns location/site of land degradation. The studied geographic location or site is defined; e.g. Brazil, Northern Africa, or Southeast Asia.

At this point, a whole range of open questions needs to be answered for constructing the operational chain by the means of the available software components. Part of this process will be done on a human-based method if no quantitative criteria can be defined at that time for allowing automating the choice; other questions will be solved by an automatic process.

3.2 Further alternatives and parameters

The following list displays the different other parameters that have to be set by users before the automated part of the learning takes place (see subsection 3.3).

3. The third choice concerns the sensor type. The sensor refers to the type of satellite acquisitions that are going to be selected and processed. The criteria to be considered for this option are: availability of data (location, period…), data quality, resolution, acquisitions' size, compatibility between acquisition channel and studied phenomenon etc. In this paper, we are interested by land degradation analysis with remote sensing data for large areas of the Earth's Surface. We, therefore, use low-resolution satellite data acquired by sensors such as NOAA or MODIS, which have a large spatial coverage.
4. According to the selected sensor, different products can be used among those available. For instance, if the sensor choice is MODIS, the available products range from: a) MODO9GQK: daily red and near infrared reflectance, 250m, quality and orbit coverage metadata, b) MODO9GST: 1km, quality metadata, (products a and b go together) c) MOD13Q1: 1) 250m, red – infrared reflectance, 16days, 2) 250m, NDVI, 16days, 3) 250m, EVI, 16days. Criteria for choosing among them are: the quality of the temporal profiles we get after pre-processing, the need for daily acquisitions or not, the availability of the data, the cost of pre-processing etc.
5. Next is the choice of pre-processing, which aims suppressing noise from the chosen data (e.g. cloud correction, apply mask of atmospheric quality, etc).
6. Another option concerns the kind of image analysis to pursue. Should the analysis be conducted for pure pixels only (see point 8 of the next subsection to understand what a pure pixel is) or for all pixels in the images? One criterion for this choice could be the ratio of pure pixels to mixed pixels in the training data.
7. Following this decision, another issue that arises is which method to use if we decide an all pixels analysis. One alternative for analyzing mixed pixels could be to consider the three main classes around it and then "un-mix" this pixel by computing the proportion of these classes within it. Another alternative is to "un-mix" the pixel with all the available. A criterion for selecting one or the other method is the homogeneity of the neighbourhood.

3.3 The automated part of the learning phase; the final parameterization process

The following paragraphs demonstrate the core choices of the learning phase, supported by automated tasks. The objective is, three-fold:

- Definition of the land use labels L_L that can be observed on low-resolution satellite images sequences;
- Definition of the characteristics (features) of the temporal profiles for each identified land use label;
- Selection and training of a classification method for the low-resolution images sequences.

The initial resources consist of at least one high-resolution satellite image acquired over the learning area[1], classified according to a high-resolution nomenclature, which corresponds to different types of land use). This high-resolution classification image constitutes a snapshot of the area and does not inform on any dynamic process such as land degradation. The second input of the learning phase is the sequence of low-resolution satellite acquisition over the area of interest (chosen in step 4), including the learning area documented by the classification.

8. This parameterization step reflects the first objective of defining a low-resolution nomenclature; the labels $L_{L,}$ associated to the low-resolution images, corresponding to different types of land use that have distinct spectral and temporal characteristics in these images. This nomenclature is usually different from the high-resolution

[1] A small part of the study area, for which high-resolution satellite images are available and a number of data acquired during field campaigns.

nomenclature L_H. To do this, the tasks that should be undertaken are:

- Finding pure pixels in the low-resolution satellite images, pure pixels contain only one high-resolution land-use label L_H.
- Computation of the temporal profile of the Normalized Difference Vegetation Index (NDVI) for each pure pixel. These curves are computed if the product (chosen in step 4) concerns red and infrared reflectance. If the product is a vegetation index, nothing further has to be done.
- Cloud-screening and temporal filtering of NDVI profiles.
- For each high-resolution label L_H testing of the profiles of all the low-resolution pixels belonging to that label to find the potential existence of any common characteristic of these profiles, so that the low-resolution land-use label L_L, defined with the L_H value, can be easily discriminated from others.
- Comparison of the temporal profiles of the remaining classes, which can result into:
  - a) Class splitting; has input the pure pixels belonging to one high-resolution class and results in two or more groups, defining different classes. This splitting is due to the heterogeneity of the set of temporal profiles of the high-resolution class. Splitting is only done on the original high-resolution classes.
  - b) Class merging; has input the pure pixels of two different classes and results into one class due to the similarity of the temporal profiles of the pixels.

The result of this process is the identification of the low-resolution labels L_L, using the high-resolution nomenclature as a guide.

9. The second issue is the selection of features on the temporal profiles for each low-resolution label. To do this, the tasks that should be undertaken are:
   - Fit of the temporal profiles by mathematical curves (polynomials), in view of filtering the remaining noise, and mostly to provide a mathematical representation of profiles from which high-level features can be computed.
   - Computation, from the algebraic expression of fit profiles, of:
     a) Features describing the whole temporal profile, e.g. number of modes, duration and growth of main growth/decrease period, amplitude of NDVI variation, mean NDVI, etc.
     b) Features per mode (if number of modes>1), each mode being described by e.g. date of maximum, amplitude, duration, etc.
   - Different methods are possible for the final selection of features, aiming to minimize intra-label variance and to maximize inter-label variance. The selection of the most adapted method is highly dependent on the statistical distribution of features and their correlations: Expectation-Maximization, Discriminant Analysis, decision trees are examples of such methods.

10. The third issue is the selection of the classification method, its training and its testing. Hence, we need to be informed on the characteristics of classification methods and to be provided with criteria for the selection. To do this, the tasks that should be undertaken are:
    - Choice of a classifier method, according to different criteria of the specific application, e.g.:

    a) Maximum likelihood if a large number of training samples is at hand, if features are uncorrelated and normally distributed (this is rarely the case in practical remote sensing problems);
    b) Decision-trees if clear cut-off values can be applied to individual features for class separation;
    c) Fuzzy-logics based methods, if some features present saturated histograms, or when some features provide information for specific classes only (and hence must not be used for other classes): the distribution of features per class is described by membership functions. Fuzzy toolboxes are at hand to model saturated as well as normally distributed features. The membership functions are finally merged using information-fusion paradigms (e.g. Dempster-Shafer combination rule).

    - Choice of a training sample for the land degradation label. This is performed by analysing change detection between two high-resolution images.
    - Training of the classifier, by analyzing the training samples to provide the required input data to the classifier. In the case when fuzzy classification has been selected, the training consists of:

    a) Identifying the relevant features for each low resolution class;
    b) Modelling the feature distribution by membership functions;
    c) Parameterizing the rule of combination for membership functions, generating class likelihood values.

    - Application of the classifier and tuning of a threshold to be applied on the computed class likelihood, to discard pixels classified with low confidence (so called non-classified pixels).
    - Test of the classification performance; produces the confusion matrix of a test area.
    - Eventually, processing of the non-classified pixels; within these pixels, the percentage of content for the different low-resolution labels is computed using a linear mixture model.

The process of developing the ontology for this first level is crucial. It supports scientists in further defining the operational chains, for different applications, by giving a clear view and a complete understanding of the choices that have to be made to address the specificities of each case.

4. PARAMETERIZATION OF THE GENERIC ONTOLOGY; STEPPING DOWN THE TREE

This section illustrates the derivation of ontologies from the top to the last level of the tree. At the top (level 0) is the generic ontology for detecting land degradation, which documents all the choices that have to be made in the context of land degradation assessment using Earth Observation means, as well as the sequence of making the choices. As demonstrated, the ontology documents the available expert knowledge for land degradation assessment cases in general. It does not provide any specific answers of what choices have to made, does not "favour" any methodologies against others, and does not exclude any possibilities. What it, nonetheless, does is to reveal and justify the whole range of choices, methodologies, and their subsequent results. The parameterization starts at lower levels of the tree, as we shall see in what follows.

At the first level, we consider different types of land degradation. For making the tree finite we consider only three; erosion, deforestation and desertification. Regarding these applications, one main difference is the speed of the degradation process, which corresponds to different data requirements. Desertification and erosion are slow and progressive processes of vegetation and soil removal, requiring analysing time-series of satellite images over several years; deforestation on the contrary is an abrupt process (forest cut or burning) only requiring a year of data, in which sudden vegetation removals are looked for. Thus, erosion, deforestation, and desertification correspond to different operational chains and associated ontologies. In this paper, we chose deforestation as the case study application for illustrating the method.

The second level reflects the choice of location/site for the application. Land degradation phenomena happen in different locations around the world, Brazil, Nigeria, South Africa, among others. Our case study is conducted for Taquari basin in Brazil[2].

At the third level, two data source types are available, MODIS or NOAA, each with its own set of parameters and values. Regarding data sources, as the spectral content and the spatial resolution (1km for NOAA, 250m for MODIS) of both sources are different, the data do not carry the same information on the occurring processes. This implies using different methodologies, as for instance sub-pixel modelling for getting spatial details despite the low spatial resolution of NOAA. MODIS data are used in the following.

This process of choosing between different options goes on for all the remaining levels of the tree reflecting the set of choices described in the previous section. When all choices are made, we are left with a complete task ontology, which corresponds to an operational chain that can be pursued by an end-user. The next section presents in detail the development of the operational chain we use as the case study.

5. THE LOWEST LEVEL; THE OPERATIONAL CHAIN

This section gives an example of the lowest tree level; the ontology of an operational chain for identifying deforestation in Brazil with MODIS data. This operational chain is implemented for detecting deforestation areas in the Taquari basin. The following subsections explain the methodology used to assess the deforestation areas.

5.1 The methodology for deforestation assessment in the Taquari basin

Daily MODIS red and near infrared reflectance at 250m resolution have been processed to monitor the Taquari area. The temporal profiles acquired by MODIS have been cloud-filtered and approximated by polynomials. Afterwards features haven been computed from these polynomials, finally the detection of deforestation has been performed in the associated vector space. An example of deforestation detection is shown on figure 2. The results can be validated with high resolution Landsat images, as illustrated in figure 3, where MODIS pixels, detected as deforested, are validated by examining the same area on Landsat images acquired before and after the detected event. The cut of the forest is clearly visible in the most recent Landsat image, and has the same shape as the detected deforested MODIS pixels.

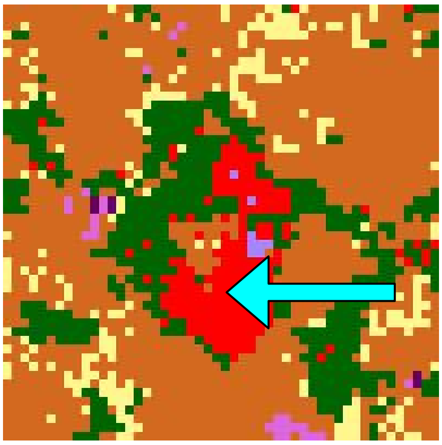

Figure 2: Example of detected deforested areas (in red).

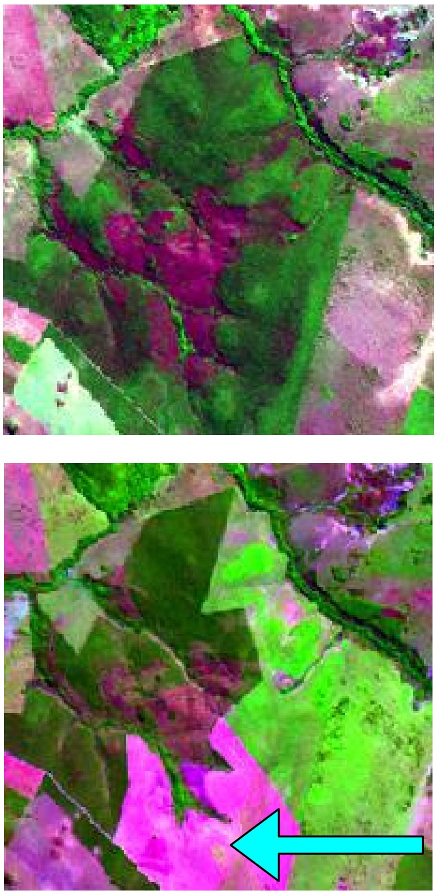

Figure 3: Validation of deforestation detection with Landsat images acquired before the deforestation process (above) and after (below).

5.2 The ontology of the operational chain

In turn, we have developed an ontology that documents and explains to the end-users the process of identifying the deforested areas from satellite images time series. It is an ontology of the operational chain; namely it documents the sequence of steps that have to be executed in order to achieve the above-described goal.

[2] The case study is partly funded by a INRIA/CNPq project

The developed ontology belongs to the class of task ontologies, it is addressed to the end-users, and it clarifies significantly the operational chain. It was created in the OWL (McGuinness & van Harmelen, 2004) language using Protégé Ontology Editor (2006).

Initially, the different tasks or steps of the operational chain were identified and, in accordance with them, the corresponding class *task* and its subclasses were created in Protégé: e.g. *build sequences*, *cloud correction*, *mask generation*, etc. Every task corresponds to a software component of the operational chain. This set of components was documented in the ontology by creating the class command and its subclasses. Several other classes were included in the ontology such as *input, output, parameter*, etc.

Similarly, the attributes of these classes were created. The most essential are:

- Has prerequisite task/is prerequisite task of (define the sequence of tasks/steps)
- Has input/is input of
- Gives output/is output of
- Is executed by command/executes (define the correspondence of each task/step to a specific command, which in turn corresponds to a software component)

For each class the necessary conditions/restrictions were determined. Namely to each class properties and relations between these were ascribed, that is why we can speak of an ontology and not of a simple taxonomic structure of the task/ steps. For making the ontology better documented and for the understanding, by the end-users, of its various classes, comments were introduced to each class.

After this process had been completed, the html files of the ontology were created for each one of the ontology elements (class, property etc). These constitute a web site, which can be given to the end-users so that they can consult it each time they follow the specific process of identifying deforestated areas. Several connections between the html files are generated automatically, when generated by Protégé; others were added manually, when believed necessary for the better understanding of the user.

6. CONCLUSIONS AND FURTHER WORK

In this paper, we presented a methodology for documenting, using ontologies, satellite images-based land degradation monitoring. We propose adopting a tree structure for organizing these ontologies since this structure allows the understanding of the interdependencies among the different possible operational chains and permits updating and refining expert knowledge.

This research has proven that the tree structure gives solutions to the question of reusability of: a) the ontologies themselves, b) the documented knowledge therein, and c) the software components run in the applications. Moreover, it has, revealed that extendibility (adding new parts of knowledge to the structure) is, through this approach, feasible.

We identify though, two future directions of research. The first involves theoretical issues of ontology development addressing the question of how generic we can be at level 0. Generic ontologies are always subject of research and are therefore, susceptible to modifications. We could get more abstract, extending the tree upwards, leaving more room for parameterisation, and reasoning about things to the next levels of the hierarchy.

The second one concerns automation of ontology retrieval from the tree. The issue, here, is to automate the process of deriving the ontologies in the lower levels from the generic one. We can also envision that this automatic derivation of ontologies will be coupled with the creation of the operational chain by linking the choice and parameterization of the software components with the choices performed within the tree.

ACKNOWLEDGMENTS

The high-resolution classification and figure 2, have been provided in the context of the INRIA-CNPq project ENVIAIR by Mrs M.S. Meireilles.

REFERENCES

Chandrasekaran, B., Josephson, J.R. and Benjamins, V.R., 1998. “Ontology of tasks and methods”. http://www.cse.ohio-state.edu/~chandra/Ontology-of-Tasks-Methods.PDF (accessed 12 Nov. 2007)

Gomez Perez, A. and Benjamins, V.R., 1999. Overview of Knowledge Sharing and Reuse Components Ontologies and Problem-Solving Methods. In Proceedings of the IJCAI-99 Workshop on Ontologies and Problem-Solving Methods (KRR5), 2 August, 1999, Stockholm, Sweden, (Amsterdam, The Netherlands: CEUR Publications), pp. 1-15.

Guarino, N., 1998. Semantic Matching: Formal Ontological Distinctions for Information Organization, Extraction and Integration. In Information Extraction: A Multidisciplinary Approach to an Emerging Information Technology, M.T. Pazienza (Ed.) (Berlin Heidelber: Springer Verlag), pp. 139-170.

McGuinness, D. L., and van Harmelen, F., (Eds.), 2004. “OWL Web Ontology Language Overview”, W3C Recommendation. http://www.w3c.org/TR/owl-guide/ (accessed 12 Nov. 2007).

Oberle, D., 2005. *Semantic Management of Middleware*. US: Springer Science, pp. 268

Protégé Ontology Editor and Knowledge Acquisition System, 2006. http://protege.stanford.edu/ (accessed 21 Apr. 2007)

Rajpathak, D., and Motta, E., 2004. An ontological formalization of the planning task. In Proceedings of the Third International Conference on Formal Ontologies in Information Systems (FOIS'04), A. Varzi and L. Vieu (Eds.) (Amsterdam: IOS Press), pp. 305-316.