Knowledge Basis and Query Tools for a Better Cumulativity in the Field of Archaeology: The Arkeotek Project

Valentine Roux, Nathalie Aussenac

Knowledge Bases and Query Tools for a Better Cumulativity in the Field of Archaeology: The Arkeotek Project

Valentine Roux[1], **Nathalie Aussenac-Gilles**[2]
[1]CNRS, UMR 7055, Nanterre, France
[2] IRIT, CNRS UMR 5505Université de Toulouse, France
valentine.roux@mae.u-paris10.fr, aussenac@irit.fr

Abstract
The Arkeotek project aims at building knowledge bases in the domain of the archaeology of techniques. These knowledge bases are made up of documents structured in data and interpretation rules, the latter being understood as inference operations performed to generate conclusions or interpretative hypotheses. Such structured documents are obtained through the logicist analysis, a 30-year old term given to an ensemble of research aiming at clarifying the mechanisms and foundations of the reasoning which organize our scientific constructs (GARDIN 2003). Nowadays, only logicism proposes an efficient methodology for extracting the reasonings contained in our scientific publications and therefore for building corpuses of inference rules. In this paper, we focus on the tools and resources designed for querying such corpuses: a domain ontology associated with a terminology, a semantic annotation tool as well a query tool. The originality of our approach is to support the corpus and domain knowledge evolution. The ultimate goal is to give the archaeologist the possibility to consult archaeological interpretations on specific subjects, as well as the foundations of these interpretations including data bases.

Keyword : knowledge bases, semantic annotation, ontology evolution

1. Introduction

*The Arkeotek Project (*www.arkeotek.org) has three complementary and interdependent aims, serving knowledge cumulativity in the field of human sciences (GARDIN *et al.,* 2004). ***The first*** one is to develop methods and tools for constituting "logicist corpuses" made of documents structured in data and interpretation rules, the latter being understood as operations of inferences leading to conclusions or interpretative hypotheses (GARDIN, 2003). Each rule consists of a proposition (conclusion) and the antecedents (premises) that support it under the form - GIVEN *i*, THEN *p* - to be read with the prudence needed for scientific work, "If it is taken as proven that… Then it can be reasonably supposed that…". ***The second aim*** is to constitute, based on this model, various corpuses in archaeology of techniques, a field of excellence in Europe. ***The third*** aim is to use semantic web technologies and an ontology to make it easy to browse these corpuses and look for specific inference rules. For this purpose, an ontology as well as an automatic annotation tool has been developed (AUSSENAC-GILLES *et al.,* 2006). Semantic annotation enables to query the logicist corpuses both on the premises and conclusions of interpretation rules.

These corpuses will have a double function: a function for guiding the researchers in scientific interpretations, and a documentary function for sharing interpretations and facts mobilized by the proposed interpretations. These corpuses should contribute directly to the cumulative process of knowledge as well as to a research dynamics. Furthermore, setting up logicist corpuses in the field of technology should serve as a model for the type of corpus which could be developed in the field of human sciences.

This paper presents the SCD format (section 2) used to represent inference rules as well as the challenge raised by our project (section 3). Then we report the ontology and annotation tool developed for information retrieval in SCD corpuses. We rely on an experiment carried out in the domain of grinding stones (section 5) to report the strengths of this project to manage, to exchange and to discuss scientific findings.

1. Collecting logicist documents: the Logicist Corpuses Project

Four steps are required to constitute a logicist corpus. First of all a) a significant number of scientific texts concerning the archaeology of techniques have to be transformed into logicist documents, that is documents presenting the scientific construct under the form of an inference tree, linking initial propositions to final propositions through successive intermediate level propositions obtained by an inference process; b) the data connected to these scientific constructs

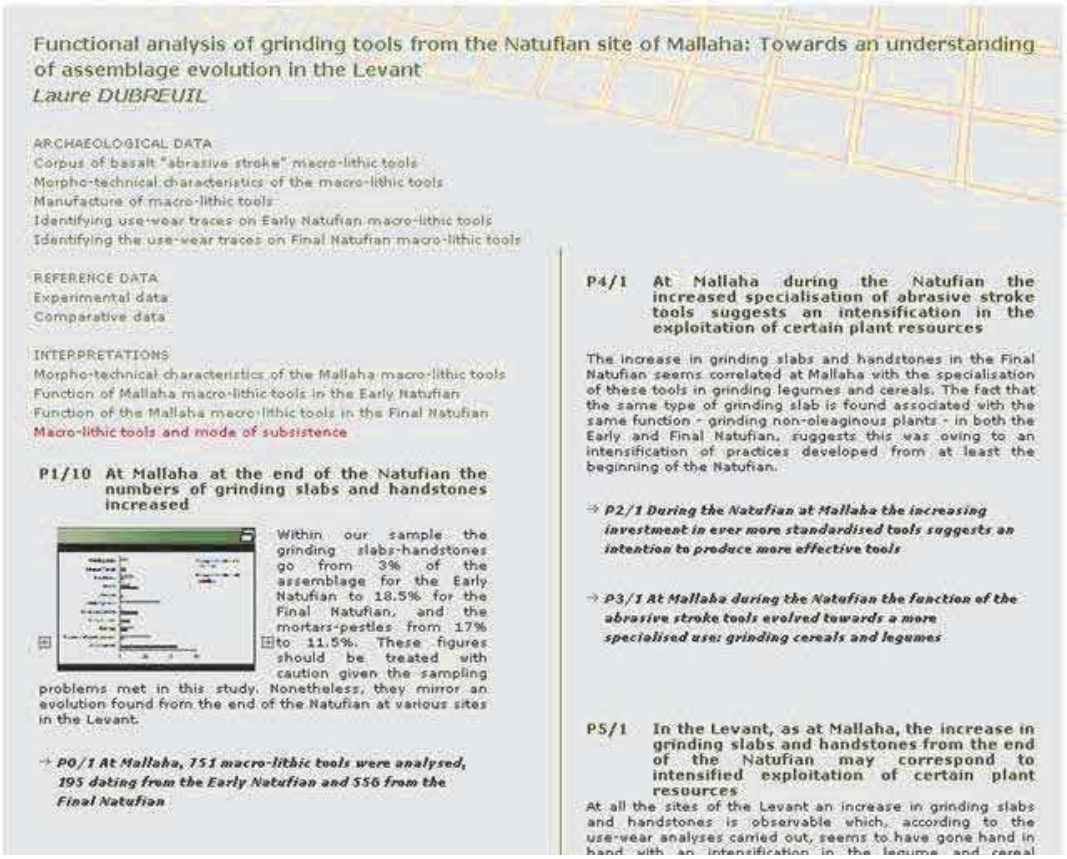

Figure 1: *Example of publication of a paper into SCD format in The Arkeotek Journal.*

have to be collected; c) these documents have to be translated in English or French (depending on the original language), d) they have to be then published in the SCD format ('Scientific Construct and Data', this format has been developed by the publisher Editions Epistèmes) in *The Arkeotek Journal (*www.thearkeotekjournal.org).

1.1 The SCD format

Briefly, the SCD format edits logicist documents and permits their rapid reading as well as an easy exploration of their constitutive elements - data bases and inferential operations -. Exploration of the constitutive elements is done according to four levels of reading, each level corresponding to a more or less detailed consultation of the scientific construct.

The first level of reading is that of a general outline listing the main blocks of propositions organizing the scientific construct. The second level corresponds to a rapid reading of the different propositions contained in each block, i.e. a rapid consultation of all the propositions that organise a scientific construction: a) the initial propositions which are generally distributed between three blocks: the observational, the comparative and the reference data, b) the interpretative propositions which include the intermediate and the final propositions, the former linking the latter to the initial propositions. The third level of consultation consists in reading: a) the commentaries developed for each proposition and written in natural language, b) the details related to the initial propositions if given and c) the antecedents upon which the interpretative proposition rest. In the latter case, the antecedents are indicated in order to unravel the logico-discursive operations founding the proposition. The fourth level of reading is that of the series of data mobilised in building initial propositions. These data are given under the form of illustrations (tables, graphics, images, videos, pdf). These illustrations accompany initial propositions or details of initial propositions.

1.2 Inference rule corpuses as supports for knowledge exchange

The SCD publishing of logicist documents enables: a) a rapid reading of the rules used by the researchers to obtain or support a result ; this is a major point since nowadays there is no automatic tool enabling to extract content of scientific texts and therefore able to propose solutions to the "crisis of scientific information"; b) easily understandable scientific reasoning, and, in return, better sharing of knowledge within the discipline; c) exhaustive access to the databases on which the scientific constructions of a field are based which is a major

advantage as compared to the printed publication which constantly have to face inherent problems linked to the restricted space available for research databases - the archaeologists should be especially sensitive to these expectations, in so far as the present publishing process does not allow experimental data to be generally shared even though these data are indispensable to the dynamic of their research -; d) the auto-archiving of research data and a solution for the perpetuation of the indexing of the data.

The SCD publishing of logicist documents enables also the constitution of knowledge bases that collect all available rules. They include the rules from the articles of the journal *The Arkeotek Journal* and from monographs of the Référentiels collection (www.arkeotek.org). They can also include rules published in other journals or books so long as these rules have been rewritten according to logicist principles (such as the rules published by Gallay in 2007). At the time being, we have a unique rule base including the few hundred inference rules found in the articles and the digests published by *The Arkeotek Journal*. These rules can be considered as either 'local' or 'universal'. 'Local' rules correspond to ordering operations (classification) or comparative operations (comparing two object collections). They are local in the sense that they apply strictly to the body of data studied in the article (or book). They are expressed under the form KNOWING *i*, THEN *p.* The rules with a 'universal' character are those rules which are not specific to the studied body of data. They call generally upon implicit reference data (the 'common sense'). Their validity can be assessed in terms of transferability. For this purpose, they have to be generalised and formulated under the form IF … THEN ... In *The Arkeotek Journal*, the inference rules contained in each article ('local' and 'universal' ones) are listed per article in the tab CORPUSES/ RULE BASES. The inference rules with a 'universal' character are given a specific access in the menu of the SCD article in the tab 'rules of inference'. These rules are put to debate through a forum. The Arkeotek editorial board makes a first comment by generalising the rules under the form IF … THEN. The scope is to assess their validity through a formulation enabling their application in various chrono-cultural situations.

In archaeology, cumulative process of knowledge involves mainly propositions obtained through 'local' inference rules, which are mainly descriptive. Propositions obtained thanks to 'universal' inference rules are discussed but rarely subject to empirical verification. Therefore they are not taken up in the cumulative process of knowledge.

2. Queries on inference rules: the DYNAMO project

2.1 Motivation for querying rule corpuses

Making logicist documents available on the web enables an easy access to these information resources. People can access to the data and the rules of inferences, debate their validity and compare the rules from one corpus to the other. Knowledge and techniques from the Semantic Web community can contribute towards the realization of cross-corpus access (HOLLINK *et al.*, 2008). The Arkeotek project data set is particularly suitable for the Semantic Web approach, since rule corpuses form rich and well-structured knowledge sources. Moreover, existing controlled vocabularies and thesauri can be used to index large collections of text or inference rules in our case. *The Arkeotek Journal* web site can be considered as a portal where semantic search is required to get precise information. Semantic annotation enriches rules with a formal representation of their content in the form of concept lists or conceptual graphs. This annotation requires to define adequate domain ontology, and to match the terms used to label concepts with the language used in the rules of inference.

Questioning the rule base implies questions both on 'local' and 'universal' inference rules, bearing either on their premises or on their conclusions. Two sorts of questions are considered, *general* and *particular*. *General questions* call upon inference rules as a guide for interpretation. Questions may be "Given *i* attributes, what can I say?" or "What attributes do I need for founding interpretation *j*". *Particular question*s call upon inference rules as a source of documentation along with the critical apparatus. Questions are for example "What are the characteristics of *i* material?" "How was organised the production of *i* material?"

In any case, answers are inference rules, the premises and/or conclusions of which match the request. The user can then consult archaeological interpretations on specific subjects, but also the foundations of the archaeological interpretations, including data bases. Cross consultation of corpuses make it possible to contrast and compare rules defined in various sub domains of archaeotechnology.

2.2 An Ontology for Semantic Annotation: principles

The ontology covers the domain of the Archaeology of Techniques, with a rich lexical component so that concepts can be used to index text: a sentence will be indexed with all the concepts which have a linguistic realization in this sentence. In this regard, it is a lightweight model with a terminological component: we call it a

termino-conceptual resource. The ontology content, its design principles and its structure are influenced by its use for textual annotation. *The ontology design principles* are the following (AUSSENAC-GILLES, 2006):

- Its scope covers the domain determined by the set of rules to be annotated but only those. Concepts are distributed between the ones related to the description of objects (intrinsic and extrinsic attributes), and the ones related to their interpretation.
- Concepts and terms are those required to adjust requests and rules. The model intends to reflect the conceptual categories that can be differentiated through the use of language.
- The ontology has been made "*a minima*", with few properties and no formal axiom, but with a rich and extended set of terms labelling each concept. The ontologist makes it evolve when rules from a new specialized technical domain are added to the collection.

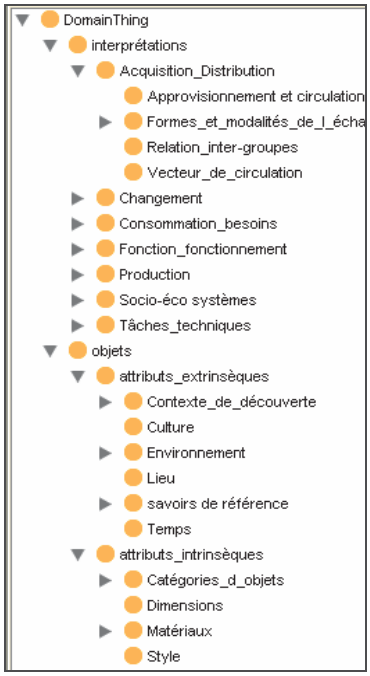

Figure 1: *The Arkeotek ontology high-level concepts.*

The ontology data–model makes explicit the concept lexicalizations using a term structure and a denotation relation from terms towards concepts (REYMONET *et al.,* 2009). At the time being, the ontology gathers concepts and terms to express technical interpretations made by archaeologists as well as those required to describe objects (extrinsic and intrinsic properties) in the domains of ceramics, lithic industries and beads. It contains 372 terms labeling about 82 concept classes with 5 properties.

Each semantic annotation of a rule can be represented as an OWL graph of term and concept instances, with semantic relations linking concept instances. The annotation process relies on mapping terms in the ontology with the words used in the rule premises or conclusion. This mapping relies on a measure that allows small spelling variations to map nouns, adjectives and verbs whatever their gender, number or conjugation when they are used in the sentence to be annotated. When setting the software parameters, the ontologist has to decide whether to take into account semantic relations or not, and which ones. Based on the concepts identified in each proposition, the system exploits the semantic relations connecting these concepts in the ontology to connect the instances in the annotation graph. In the particular case of the Arkeotek corpus, annotations do not include semantics relations between concepts. To sum up, the annotation algorithm automatically generates an OWL graph for each proposition, and the ontologist or corpus editor checks, improves or modifies it before he validates this representation.

2.3 Ontology and annotation tool

Querying the rule base implies first to develop a relevant ontology, then to annotate each rule, and then to express and match queries. The tools and infrastructure for all these tasks have been defined within the DynamO project (http://www.irit.fr/dynamo/). The originality of this project is to define a unique environment (TextViz, a plug-in of the Protégé ontology editor (http://protege.stanford.edu) for ontology management and use for semantic annotation, in order to anticipate ontology evolution consequent to the corpus evolution and changes in uses needs. As long as new inference rules are added to a corpus, the ontology is adapted by adding new terms or concepts, or more deeply modified, so that new rules could be precisely indexed. The Dynamo project experiments an extension of the notion of ontology, a termino-ontological resource that enriches an ontology with terms denoting each concept (REYMONET *et al.,* 2009). Terms play the role of linguistic markers to identify that concepts are mentioned in rule propositions.

Another original feature of the Dynamo project is to provide support for ontology evolution in the TextViz system: the ontology is modified every time new terms or concepts are required for rule annotation. The Dynamo project decided to test and compare two different and complementary approaches for ontology evolution: (1) a supervised evolution process based on the ontologist's initiative (REYMONET *et al.,* 2009) – the ontologist can manually select a phrase in a sentence and define it as a term or as a concept label connected to this term; (2) an adaptive multi-agent system that makes evolution suggestions on the basis of the terms and semantic relations extracted from the corpus with NLP tools (SELLAMI *et al.,* 2009).

With TextViz, the evolution cycle is a loop where the evaluation of quality criteria (like a list of concepts that have to be identified in each

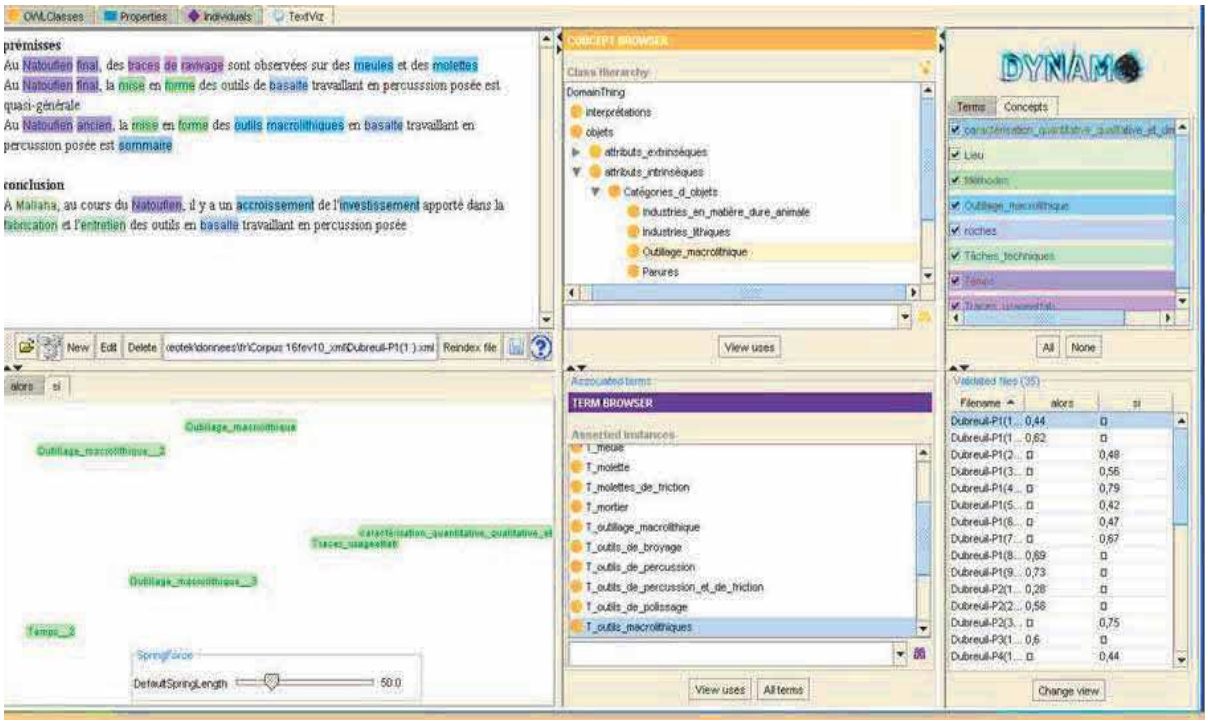

Figure 2: *Proposition annotation with TextViz. On the left top, an inference rule made up of premises and a conclusion.Just below it, a graphic view of its annotation. In the middle, top: the ontology concept hierarchy, bottom: terms labelling the "outillage_macrolithique" concept. On the right, top: annotating concepts of the rule indicated with colour, bottom: inference rules sorted according to their validation score.*

annotation) may lead to modify the ontology, and where changes in the ontology may lead to launch a new annotation process. The interface displays all the resources and information required to support this cycle: the hierarchy of concepts with, for each of them, the terms that denote them; the current quality of the annotations of the rule corpus and their validity; three views on how a selected rule is annotated with concepts: a graphic view, a list of concepts and a textual view where tagged phrases are high-lighted with colors.

2.4 The search and editorial tools

Querying the rule base also requires a set of tools for the end-user, including an adequate browsing and querying interface. Graph-based semantic annotation makes it possible to express simple queries and match them with appropriate documents (HILDEBRAND, 2008). This author promotes the combination of various vocabularies to guide the formulation of queries with precise words. The Arkeotek overall interface should make it possible to browse the rule collection paper by paper or within a given sub-domain, or to query the rule corpus with natural language requests. The ontology can also be use as a support to express requests and to browse the rule collection.

The rule and data reading interface will comprise a) a natural language query device, b) an ontology browsing device that helps build request with concepts, c) a rule reading device to browse any request answers. Browsing the rules returned as answers makes it possible to compare or contrast these rules. It offers an innovative reading of the different rules with one or several corpuses, and the possibility to consult their premises and conclusion. Thereby, the reader can check the strength and validity of the available rules and data by consulting back the original publications.

3. Experiment: Testing the "grinding material" corpus

The Arkeotek Journal has published three articles and one digest on grinding material (different bodies of data, different chrono-cultural periods and area). This thematic corpus is made up of 83 inference rules. Out of these 83 inference rules, 14 rules can be reformulated as so-called 'universal' rules in the forum.

The corpus has been presented and discussed within a workshop which gathered domain experts (http://www.arkeotek.org/index.php?option=com_content&task=blogcategory&id=31&Itemid=43). Two categories of questions have been raised, technical and theoretical. On the technical side, the Arkeotek Project has been acknowledged positively: the logicist documents and the search tools proved to be very efficient for documentary purpose (rules and data) as well as for comparing inference rules. For example, queries have been formulated about

standardised grinding material: “what can I say if I have standardised grinding tools?” The returned rules of inference, proposed by two authors using different terminologies, propose that standardized grinding tools could be interpreted as efficient tools. By comparing the rules of inference, it appeared a) that terminology and theoretical framework can be different but the rules of inference very much comparable, b) that rules of inference can be very current in one field but however not well founded; indeed, discussions highlighted the fact that standardized tools do not express efficiency but specialisation of the manufacturers. In other words, the query tool enabled us to highlight the fact that most of our interpretative rules are implicit and therefore not really discussed, therefore preventing from a proper cumul of knowledge. On the theoretical side, the Arkeotek project has been perceived as a project which should impact on our researches: the logicist documents show that most of our constructs are made up of “middle range” propositions, as well as of quite useless propositions (a significant number are in fact redundant); the ones calling upon ‘universal’ rules appear, on the contrary, the most interesting ones since they enable us to obtain “high level” interpretations. Nevertheless the assessment of their foundations needs more experiments. In this regard their formulation under the form of rules opens paths to new researches.

4. Conclusion

In conclusion, *the Arkeotek Project* is definitely acknowledged as efficient in terms of extracting knowledge from linear texts, providing formatted documents enabling the development of searching tools, and offering a library of scientific inference rules along data and a solid critic apparatus (since the premises of each conclusion are explicit). In this regard the Arkeotek Project provides the means for a better research dynamics. By extending its network among the scientific community, it should be slowly appropriated by researchers in Human Sciences as a powerful technical and epistemological tool for constructing and disseminating scientific results.

The *Arkeotek project* illustrates the importance of structured documents for developing tools which enable us not only documentary search, but also scientific search. The current experiment in the domain of grinding material and the use of the TextViz tool confirmed the gain brought by semantic web technologies. This tool is also experimented in two other domains like car electronic-fault diagnosis and repair or software maintenance. To be relevant, the TextViz must offer good evolution capabilities that will keep the termino-ontological resource up-to-date with regards to the corpus and the domain knowledge.

Bibliography

AUSSENAC-GILLES N., ROUX V., BLASCO P., 2006, The Arkeotek project: structuring scientific reasoning and documents to manage scientific knowledge. *Proc. of the workshop on Indexing and Knowledge in Human Sciences*, Nantes (F), Proceed. Of Semaine de la connaissance, June 2006. (on line : http://www.irit.fr/SDC2006/)

GALLAY A. 2007, 73 propositions pour rendre compte des sociétés alpines et pré-alpines du IIIe millénaire avant J.-C., *In* J. Guilaine (ed.), *Le Chalcolithique et la construction des inégalités.* Tome I. Le continent européen. Paris : éditions Errance, Séminaire du Collège de France, p.93-123.

GARDIN J.-C. 2003. Archaeological Discourse, Conceptual Modelling and Digitalisation: an Interim Report of the Logicist Program. *In* Doerr M. & Sarris A. (eds), *CAA 2002 - The Digital Heritage of Archaeology,* Hellenic Ministry of Culture, Archive of Monuments and Publications, Heraklion, p. 5-12.

GARDIN J-C. and ROUX V. 2004. The Arkeotek project : a european network of knowledge bases in the archaeology of techniques. *Archeologia e Calcolatori*, 15, p. 25-40.

HILDEBRAND M. 2008, Interactive Exploration of Heterogeneous Cultural Heritage Collections. In A. Sheth et al. (Eds.): *ISWC 2008,* LNCS 5318, Springer-Verlag Berlin, p. 914–919.

HOLLINK L., ISAAC A., MALAISÉ V., SCHREIBER G., 2008, Semantic Web Opportunities for Digital Libraries. *in 32nd Library Systems Seminar of the European Library Automation Group* (ELAG 2008). Wageningen, The Netherlands.

MIRZAEE V., HAMIDZADEH B., IVERSON L., 2005, Managing Change in Ontologies, in *Proceedings of the IEEE International Conference on information Reuse and Integration (IR 2005),* Las Vegas, Nevada, USA, p. 253-258.

REYMONET A., THOMAS J., AUSSENAC-GILLES N., 2009, Ontology Based Information Retrieval: an application to automotive diagnosis. *International Workshop on Principles of Diagnosis* (DX 2009), Stockholm, M. Nyberg, E. Frisk, M. Krisander, J. Aslund (Eds.), Linköping University, Institut of Technology, p. 9-14.

SELLAMI Z., GLEIZES M.-P., AUSSENAC-GILLES N., ROUGEMAILLE S., 2009, Dynamic ontology co-construction based on adaptive multi-agent technology. *Int. Conf. on Knowledge Engineering and Ontology Development (KEOD 2009)*, Madeira (Portugal), Dietz J. (Eds.), INSTICC, p. 56-63.