

Report on “Exploitation de graphes conceptuels et de documents structurés et hypertextes pour l’acquisition de connaissances et la recherche d’informations”, par Phillippe Martin

Prof. dr Joost Breuker*
University of Amsterdam

22 septembre 1996

1 General Evaluation

The thesis of Mr Phillippe Martin shows solid and independent state of the art research in artificial intelligence and software engineering. One of the results is a new set of tools for knowlegde acquisition and information serving, but the thesis is important and of a very good, and in many respects even excellent quality, beyond the construction of these tools proper.

This qualification is motivated by the following reasons:

- The work covers a very large area of research in information and knowledge systems. The range is very broad: knowledge acquisition and its methods, techniques and tools; ontological analysis, as both practiced in the field of (computational) linguistics and knowledge acquisition; knowledge representation, in particular conceptual graph (CG) representation; information technology, and in particular, the use of hypertext.
- The work is not only broad, but also state of the art. For each of these areas, Mr Martin shows that he is a knowledgeable researcher who contributes or may contribute to current problems and research. It should be noted that Mr Martin can cover all these areas because he is capable of focussing exactly on work and issues that are relevant to his research goals. This implies that he must know his way quite well in these areas.
- The work shows a fine balance between pragmatistical solutions and excellent theoretical insights and argumentation. This balance is not only implied

*Je m’excuse de pas écrire en français, mais ma maîtrise active de l’orthographe et grammaire, et du vocabulaire précis peut avoir une influence negative sur le clarté de mon rapport.

by the fact that CGKAT is theoretically well founded, but in particular by the high level ontology, whose construction requires on one hand insight in age old philosophical problems, and on the other hand requires some pragmatic choice of the least problematic and still consistent etc. solution.

- The tools and theoretical developments are highly original. They are original in the way already existing tools and models are combined into new functionality, respectively new structures. Therefore, the work shows highly efficient (re)use of state of the art accomplishments to obtain innovative tools and conceptualizations. CGKAT uses and combines Thot, CoGITo and WordNet; the construction of the high-level ontology provides an integration of work of the most important existing models in this area.

The thesis is well written and organized. The author is a very effective and economical writer. There is hardly any redundancy, no argument is too detailed or too superficial, and there is an open-minded treatment of pro's and contra's, which makes the thesis convincing without rhetoric (in the connotative sense of the word). There is ample and detailed documentation in the Annexes, both of results of own work, and of material used from external sources.

The extensiveness and thoroughness of the work presented in the thesis may explain why there is no reporting of actual use and application of GCKAT. There is a good comparison with related, actual work, but that cannot hide the fact that CGKAT is unique in its kind, so that one is interested in how this new functionality in supporting knowledge acquisition and information retrieval is practically used.

In conclusion: there is no doubt that the thesis of Mr. Martin is of a high quality and worth defending.

2 What's in the thesis

The thesis consists of an introduction that states the objectives and approach of the work presented, after a sketch of the context of the research. This context is knowledge acquisition and information retrieval: usually these areas are viewed as typically knowledge respectively software engineering, but the author shows how they are related, i.e. how knowledge representation is the support *par excellence* for information retrieval, and how knowledge acquisition is not possible without appropriate information management tools. This view is consistently embodied in CGKAT.

The body of the thesis consist of two parts. The first part describes the state of the art in the various areas; the second part contains a report of the research performed.

- Chapter 2 discusses current views and research problems in knowledge acquisition. Knowledge acquisition (KA) is viewed as modelling both the environment as well as the knowledge system (KBS) to be implemented.

The distinctive major types of knowledge that make up the expertise to be embodied by the KBS are presented, following largely , but not exclusively the (Common)KADS methodology. The role and importance of ontologies for describing these types of knowledge takes a central position, which explains also in which context the use of CGKAT’s pre-defined ontologies have to be veiwed. By hindsight – and not mentioned in the (conclusions) of the thesis itself – one can see that CGKAT certainly has the richest and most elaborate ontologies of all KA-tools currently available. Activities required for building KBS are described following van Heijst’s (96) description. The important point here are not the activities *per se*, but the tools available to support these activities. The section devoted to the representation and organization of (the various) types of knowledge (2.3.) tries to integrate many, rather heterogeneous issues, ranging from the languages used to the nature of views. There is not a clear distinction between knowledge representation languages, languages for specifying ontologies, and languages for specifying KBS in general. Indeed, some languages are multipurpose (e.g. CommonKADS CML), but the distinction is relevant because the requirements are somewhat different. The last section describes the current topics in KA on the construction and reuse of ontologies. Various types (levels) of ontologies and their form (structuring) as well as their formal underpinning are presented in a coherent way. In summary: the chapter gives a comprehensive, up to date and clear view on theoretical issues in KA, and some aspects of tool support. Although it follows the framework prepared by other researchers (e.g. van Heijst et al, 96) and CommonKADS, it expresses largely independent and mature views of the author. Only the section on encapsulation and views, which is highly interesting, looks somewhat incoherent.

- Chapter 3 provides an overview of developments with respect to information systems. As this is not typically my domain of expertise, my comments can be less directly related to the state of the art in this area. The presentation of the various structuring and retrieval methods/approaches is very well organized and clear. Moreover, this review is directly related to CGKAT. After an overview of these approaches, including structured documents and hypertext, the concluding section presents a list of six criteria (dimensions) which appear of importance to assess and characterize information systems. In this section the author shows for five of the six dimensions what features are covered by CGKAT, in particular by combining the information management system Thot with the graphical (re)presentation tools of CoGITo.
- Chapter 4 contains an overview of the Conceptual Graph (CG) formalism as conceived originally by Sowa (1984). Although intended as a language with a formal underpinning that allows a “natural” expression (i.e. logically sound without the nasty appearances of raw logic), CG has become in AI a language for representing knowledge, and in particular for representations that support automatic natural language processing. Thus far

the use of CG for specifying or for building knowledge systems has been rather limited, and one of the important contributions of Mr Martin's thesis (and related publications) is to show how CG may be a worthwhile complement or support, if not alternative, for the languages used in KA/KE. In fact, the latter possibility is only incidentally worked out, where the author compares CG with terminological KR. Where it is clear from the chapter, which contains an excellent description of CG (sec. 2 and 3), that CG provides a very natural transition between texts that are an input for KA, and a(n operational) specification, the author (and with him many other researchers in KA) does not take into account that the requirements for representation and specification are often conflicting. For instance, for specifying ontologies in general, the commitments should be minimal. Specific languages are to be preferred for specific purposes, and the fact that general purpose languages can be "extended", or simply used to express a specific language (e.g. CommonKADS CML or DESIRE in CG) is only an example of the fact that the more commitments are built into the language, the more specific the language can be made. In fact, in CGKAT and in the work reported CG is only an operational specification language in which basic ontologies are expressed that provide the building stones for the knowledge modelling job; the CG provides a nice (also graphical) interface between these building materials and the natural language expressions. However, it may be equally true that CG is not the language for building libraries of ontologies (which require minimal commitments), or for implementing systems (computational efficiency). In summary, I think that CG is a good choice with properties that have been too long overlooked in KA, and these properties are nicely exploited in CGKAT, but – as the author also states – much further research is needed to assess its appropriateness for other roles in KA/KE.

- Chapter 5 presents the ontologies developed to support the specification of knowledge systems. It contains highly original work and the author shows a good understanding of what the issues are in this newly "emerging discipline of ontology specification". This is rare, as this discipline is marked by wild claims and rather divergent opinions of what in fact an ontology is and what it is good for. The work is not only original, but also daring. KA in any specific domain may be supported with general terminological tools. WordNet provides a huge basis, but is rather superficial for the purpose of KA, in particular where it concerns the high level ontological conceptualisations. These high level ontologies that will reoccur – often as (implicit!) commitments – in almost any domain require a more coherent and worked out top ontology. These top ontologies have in fact been the study of ages of philosophy, linguistics and only recently AI. Therefore the subject should be entered with all warnings possible and is not supposed to be a topic of a thesis. However, the work presented is of high quality as it combines in a very well argued way work of other researchers in a coherent combination that to my opinion surpasses the work of each of

these individual works. In fact, it may still take quite a number of theses to verify its full consistency and implications, and there will be certainly many revisions, but the work itself is sufficiently comprehensive to warrant such research. To me, this contribution may have far more practical and research implications than CGKAT itself.¹ Therefore I highly recommend these results to be reported in the international literature. Although some important categories are not fully worked out (verbs), the ontologie covers more than only WordNet, as it also incorporates (CommonKADS) task and problem solving ontologies.

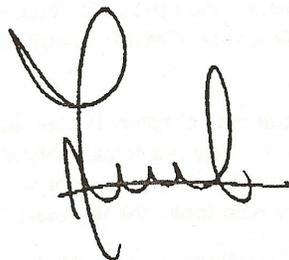
- Chapter 6, about structured documents for organising and retrieving information *and* knowledge, contains the specifications (and design decisions) of CGKAT. It reuses two already available tools. Thot for the editing of structured documents, and CoGITo for handling and graphical expressing CGs. CoGITo is rather a complex user interface, while the document elements (ED) and the CG belong to the Thot editor. This combination supports both the KA from documents into knowledge bases (KB) of CG. Vice versa, these KB may help to structure information in the documents. This is accomplished on the basis of the domain specific KB, e.g. as a process of refined, model driven KA, and/or the general knowledge implied by WordNet and the high level ontologies. This two way functionality is described in full detail, and shows the careful way in which in particular Thot is exploited to create CGKAT specific editors. The way from text (ED) to CG is via traditional KA; the way back is accomplished by the fact that for each (unique) ED there is a (unique) CG (but CGs may also express relations between CG, i.e. EDs). The chapter contains a description how the Thot editor is adapted and ‘filled’ to achieve this functionality. Not only the way CGKATS works is specified, but also how it is to be used in KA, in both a data and model driven way. CGKAT is well documented and argued for: the many screendumps help to convey some of the look and feel, but this cannot fully replace (reports of) actual experience.
- Chapter 7 describes the full architecture of CGKAT, i.e. by adding and interfacing GoGITo to the adapted Thot editor, the full functionality of CGKAT is obtained. Therefore, in this chapter the use of CoGITo, supported by WordNet and the high level ontologies, is described. CoGITo manipulates both types and (therefore typed) GC expressions.
- Chapter 8 is the concluding chapter. It contains a well structured summary of the functionality of CGKAT (and is a remarkably redundant peice of text for those who have read the chapters: however, I cannot see any more how informative it is to the occasional reader who looks for summarizing material). The real conclusions are phrased in terms of comparisons to hypertext based information systems and (other) KA tools and

¹If only because it will be relatively easy to “copy” the ontologies to other KA tools.

workbenches. The comparison with the hypertext systems is somewhat superficial, because functions and features are compared where one would expect rather marked differences in conceptualisation. However, this is rather difficult as CGKAT is largely based upon Thot, and therefore the comparison would be mainly between Thot and other information management systems, which is not directly a result of the work reported in the thesis. Therefore, the comparison with KA tools is more important. However, this comparison is also rather superficial, as it mainly concerns the functionality of tools or workbenches. The author mainly relies here on the rather elaborate comparisons of van Heijst et al (1996), rather than own views. Except for some commercially available tools, most tools and workbenches have been developed as parts of research projects, sometimes of rather limited size, to show a “a point”. As they are not intended to cover the full KA process in all its varieties and preferences, it is to be expected that the functionalities of most of these tools are rather limited. The important, rather original point in CGKAT is the knowledge representation driven information analysis and retrieval, which is only implicit and rudimentary in other KA tools. The conclusions should rather focus on this point and on the theoretical contributions of the work.

3 Conclusions

The thesis and the work reported is of a high technical and scientific *international* quality. It shows a mature balance between theoretical reflection and practical engineering, and it will be a pleasure to see it defended. The thesis moreover covers many areas in AI (natural language processing, knowledge representation and knowledge acquisition) and software engineering (information systems, in particular hypertext based ones). This is operationalised by the CGKAT tools, but to my opinion, the high level (domain) ontologies are as important a contribution, from which in particular the KA/AI research community should benefit when it is published in one or more articles in international journals. The quality and amount of work has been great. Therefore, one may regret that the thesis does not report actual experiences in using CGKAT in earnest.

A handwritten signature in black ink, appearing to be 'Paul', written in a cursive style. The signature is centered on the page.