

Research Project: ScalIntegr

Designing and Supporting Scalable Ways to Represent, Share, Integrate and Compare Precise Information

Dr Philippe MARTIN

Extended abstract

General goal, approach and guideline. The general aim of ScalIntegr, the long-term project that I propose, is to investigate and develop models, techniques and tools that support *scalable* manual and automatic ways of retrieving, extracting, representing, *integrating*, sharing, comparing, and managing information from various sources (documents, knowledge bases, knowledge providers), in a *precise, normalized and semantically well interconnected* manner. To do so, this project should refine my work aimed to support the collaborative building of shared KBs approach and extend it to integrate more distributed and/or automatic approaches to knowledge creation. However, for scalability purposes, my guideline for all the research tasks of this project will still be the same: trying to reduce the **implicit** redundancies or inconsistencies between the knowledge objects (e.g., categories or statements) and, more generally, increase the semantic organization of these objects.

Themes of research. This project will focus on finding and supporting ways to guide or automatize

- the replication of knowledge between public and/or personal KB servers - and hence the integration of knowledge from a server into another - in such a way that it does not matter which KB a user or agent chooses to query or update first (this permits to combine the advantages provided by a distributed knowledge development approach with the advantages provided by a shared KB development approach);
- the extraction and representation of knowledge (especially certain important kinds of semantic and argumentation relationships, and especially from English sentences) according to my integration and extension of important ontologies, methodologies and guidelines,
- the normalization, integration and organization of knowledge to ease its automatic comparison and hence its retrieval or sharing;
- the collaboration between users via knowledge editing protocols and knowledge content/creator valuation mechanisms - within a shared KB or a peer-to-peer network (semantic grid),
- the personalization and organization of query results; and
- the application of the above techniques to support collaboratively-built semi-formal repositories such as corporate memories and, in particular, a collaboratively-built semi-formal state of the art of knowledge engineering.

Collaborations and applications. This project, which is a broadening of my research work since 2000, fits well in the research programs of both the **KEWI** (Knowledge Engineering and Web Intelligence) team of the **I3S** laboratory and the **IC3** (Ingénierie de la Connaissance, de la Cognition et de la Coopération) team of the **IRIT** laboratory. It would also support shared research and tools with related teams at the I3S laboratory (in particular **RAINBOW** but also RL, OASIS, AOSTE, Bioinfo and MC3) and IRIT laboratory (SMAC, SIG, ILPL, LiLaC and RPDMP). Other possible applications will be in software engineering and bioinformatics, and will involve semantic grids and peer-to-peer networks. Most applications will be related to knowledge base sharing, the Semantic Web and corporate memories.

Table of Contents

1. Introduction
2. Supporting scalable ways of sharing knowledge between knowledge bases
3. Methodological and ontological models for knowledge representation and sharing
4. Knowledge presentation
5. Semi-automatic ontology integration and knowledge extraction from English sentences
6. Main test applications: collaboratively-built semi-formal repositories
7. References

1. Introduction

The general goals of this project are shared by many research projects, including major ones such as [Knowledge Web](#) and [NeOn](#) (Networked Ontologies) which are funded by the European Commission to make the Semantic Web vision a reality. However, these projects mainly focus on techniques aimed to support a "semi-independent knowledge representation approach". Indeed, most projects prefer to focus on exploiting already existing resources and adopt the current main paradigm and languages advocated by the W3C for the Semantic Web (Berners-Lee et al., 2001) (Shadbolt et al., 2006), even though many of their basic flaws are well discussed, as in (Kalfoglou et al., 2004) and (Patel-Schneider, 2005). It can also be argued that, at least nowadays, the semi-independent approach is *socially* necessary for scalable knowledge sharing and retrieval. Furthermore, genuinely supporting the "cooperative building of large knowledge bases" (or in other words, the "shared KB approach") is an ambitious long term goal that requires the design and implementation of many techniques. At a general level, the originality of ScalIntegr is that it will (i) extend my work on the shared KB approach since it is *technically* necessary for scalable knowledge sharing and retrieval, and (ii) extend it to better integrate the semi-independent approach. Extensions to semi-automatic knowledge extraction and merging techniques will also be made; such extensions are needed for both approaches. The particular approaches proposed below are also original. Each of the tasks that I propose should be done in collaboration with the team members of KEWI or IC3.

This project investigates techniques to support an efficient avenue for the incremental building of a "Knowledge Web" or genuine "Semantic Web" (on the Internet, within an intranet, or in a peer-to-peer network), that is, a *well organized* semantic network permitting *objects* (e.g., conceptual categories, beliefs, definitions, ideas, tasks, data structures, persons, organizations and products) to be accessed and compared precisely and efficiently. Each of such networks may be distributed, that is, it may span several KBs but ideally these KBs should be well interconnected and hence should not have implicit redundancies or inconsistencies. This is why I refer to each of such networks, as one global "virtual" well-organized KB.

The current nonexistence of such a network is the cause of the well acknowledged problem of information overload or difficulty of finding precise and relevant information in intranets or the Internet. For example, questions such as "What are the arguments and objections for the use of an XML-based format for the exchange of knowledge representations?", "What are all the tasks that should be done in software engineering according to the various existing 'traditional system development life cycle' models?" and "What are the characteristics of the various theories and implemented parsers related to Functional Dependency Grammar and how do these theories and parsers respectively compare to each other?" cannot be adequately answered by giving a list of documents or an unorganized list of all sentences related to these questions. They require presenting and allowing the browsing of a semantic network: (i) for the first question, a network with argumentation, objection and specialization relations, (ii) for the second question, a subtask hierarchy of all the advised tasks, and (iii) for the third question, a network with specialization relations between the various objects or attributes related to the theories and parsers.

My research report (the "report on past research works" that is included in this application) introduced a basis for the dream of Buckingham et al. (2007) - creating an Internet based infrastructure supporting a more effective dissemination, debate, and analysis of ideas than the current system of article publication - and the related dream of Hillis (2004) - a "Knowledge Web" to which teachers and researchers could add "isolated

ideas" and "single explanations" at the right place, with "mechanisms for credit assignment, usage tracking, and annotation that the Web lacks" thus supporting a much better re-use and evaluation of the work of a researcher than the current system of article publishing and reviewing. To support these dreams and the closely related goal of cooperatively-built virtual well-organized KBs, research should be undertaken on various points related to knowledge representation, organization, presentation, extraction and integration. For all of them, *the general guideline of ScalIntegr project is that ideally there should be no implicit redundancies between objects*. Some alternative formulations are: (i) the objects should be represented as precisely and uniformly as possible, (ii) they should be as small and explicitly interconnected by semantic relations as possible, at least identity and specialization relations, and (iii) the bigger and more organized the KBs, the easier it is for a software to align/merge these KBs or guide the users in entering precise and re-usable knowledge. These formulations can be seen as my work hypothesis. The designed methods and tools will be tested and refined in the context of applications such as the cooperative building of semantically-well-organized states of the art. The implementations will re-use an integration of the WebKB-2 server (Martin & Eklund, 2001) with the tools of the KEWI or IC3 teams.

2. Supporting scalable ways of sharing knowledge between knowledge bases

There are many automatic techniques and tools to extract knowledge from texts/databases or align conceptual categories from different KBs. They are useful but insufficient for precision-oriented knowledge retrieval and sharing. For example, Euzenat & Shvaiko (2007) give an evaluation of ontology alignment techniques and conclude that these tools are quite understandably imperfect although sufficient for certain applications. They recognize the need for the approach I advocate: the design of (semi-)formal KBs letting people and software agents directly exploit and save new knowledge or object alignments, that is, query, complement, annotate and evaluate the existing objects, guided by these large and well-organized KBs.

To support scalable ways of sharing knowledge between KBs (and hence the creation of a global virtual well-organized KB), my general guideline has two implications. A first one is that current knowledge extraction and alignment techniques should be combined and refined according to methodological and normalization guidelines (those I gathered so far plus new ones). This task, which is described in Section 5, will be guided by the already integrated ontologies in the destination KB. Given the guidelines, each integration of each ontology should not lead to a loss of information; this aspect will permit to further automatise the integration of newer versions of that ontology and the re-application of corrections made to previous versions of that ontology when such corrections are still necessary.

A second implication is that it should not matter which (non-virtual) KB an intranet/Internet user or agent chooses to query or update first. Hence, a research question is "how to support the replication of knowledge between KBs, whether they are publicly accessible, within intranets or within a peer-to-peer network?". More precisely, two constraints are: 1) object additions/updates made in one KB should be replicated into all the other KBs that have a scope (domain of interest) which covers the new objects, and 2) a query for which the content of a KB will not yield a complete answer (with respect to the content of the virtual global KB) should be forwarded to the appropriate KBs. Point 1 implies that two KBs having shared sub-scopes should have the same objects for these scopes. The second point implies that, in each KB, when an object within the scope of a KB is related to an object outside the scope of the KB, a reference to this last object in another KB should be used.

To satisfy those two constraints, ScalIntegr will start with an approach described in (Martin & Eboueya, 2008) and summarized in Table 1. More precisely, I propose to refine and implement this approach in two contexts: (i) for KB servers on the Web, and (ii) for KB servers in peer-to-peer networks (e.g., for a semantic grid) with the assumption that each user has its own KB server. The main difference is that a peer-to-peer network will permit to implement systematic push/pull mechanisms instead of relying on KB servers to regularly check the KBs of other servers and integrate new additions. This also means that in a peer-to-peer network the intra-KB collaboration protocols can also be integrated in inter-KB collaboration protocols. To

do so, the current protocols of WebKB-2 (Martin et al., 2005) should first be combined with those of Co4 (Euzenat, 1996). I'll undertake this research in collaboration with the members of the KEWI team working on the research theme "knowledge sharing and exchange". The work related to peer-to-peer networks will be done in collaboration with A.Pr. Eboueya (L3I, University of La Rochelle) with whom I already collaborate (as shown by my list of publications) and who uses semantic grids for e-learning purposes.

The approach of Table 1 seems the simplest approach. Indeed, (i) the approaches used in distributed databases would not work since KBs do not have any fixed conceptual schema (they are composed of large, explicit and dynamically modifiable conceptual schemas), and (ii) a fine-grained classification (via an ontology) of all the objects of the KBs is necessary for precision-oriented information retrieval: indexing KB servers according to topics or domains is a far too coarse indexation to retrieve knowledge from distributed servers (e.g., knowledge about "neurons" or "hands" can be relevant to many domains).

To the best of my knowledge there is no other project addressing the above two constraints. The works dealing with "Ontology Evolution in Collaborative Environments", e.g., (Vrandeic et al., 2005) and (Noy & al, 2006) - or (Rousset, 2004) and (De Roure et al., 2005) in a peer-to-peer or semantic grid context - are solely about accepting/rejecting and integrating changes made in other KBs, they are not about making these KBs have an equivalent content for their shared sub-scopes.

Table 1: the underlying idea of my approach to replicate knowledge between KB servers

The scope of each server is defined by a *reference collection* of objects (conceptual categories or statements) and an optional maximum depth for the chain of relations from these objects (for a totally general server, this collection is reduced to "Thing", the most general conceptual category imaginable). Each server periodically checks more general servers, competing servers and slightly more specialized servers, and

- 1) integrates (and hence mirrors) all the objects generalizing the objects defined in the reference collection,
- 2) integrates all the objects (and associated direct relations) more specialized than those in the reference collection until it reaches a maximum specialization depth if one has been specified (if so, the URL of the object is stored instead of the object),
- 3) also stores the URLs of the direct specializations of the generalizations of the objects in the reference collection (this is needed for all objects in the global virtual network to be directly or indirectly referred to).

3. Methodological, ontological and cooperation models and languages for knowledge representation and sharing

My research report quickly presented my creation/integration and enhancement of (i) knowledge representation methodologies, best practices and normalization procedures (Martin, 1993, 1996, 2000), (ii) ontologies (Martin, 2003) (Martin & Eboueya, 2007), and (iii) knowledge-based collaboration and valuation protocols (Martin et al., 2005). It also presented their integration into expressive but high-level languages (For-Links, Frame Conceptual Graphs and Formalized English) (Martin, 2002) or into the shared ontology, as with my representation of KADS models and my additions to them for explanatory information (Martin, 1996). The **refinement of my previous research on these points** is necessary for supporting the creation and extension of well organized semantic networks.

These points fit perfectly in some research themes of both KEWI and IC3: the themes "Models of knowledge and texts" and "Models and plans" of IC3, and the themes "Knowledge modelling and inferencing" and "knowledge sharing and exchange" of KEWI. The rest of this section shows this with respect to the goals of the last theme: the abstraction of mechanisms shared by different Semantic Web languages, the definition of a generic core language allowing extensions from all languages based on it, the design of a generic graph-based knowledge representation platform, the design of KB validation mechanisms based on semantic constraints, and the optimal organization of ontologies.

- My past research works and their implementation in WebKB-2 have the necessary characteristics to be a basis for such goals. WebKB-2 has a **data model** with all the basic features needed for knowledge representation (it is as expressive as KIF, the Knowledge Interchange Format, and hence its more restricted successor, Common Logics) but is graph-based (like the Semantic Web languages) and high-level, that is, it has many high-level constructs such as numerical quantifiers, time related data structures, source/creation related information on every object and other shortcuts for the representation of meta-statements and collections (Martin, 2002). This permits to use various graph-matching operators for various purposes: logic deduction, analogy, information retrieval, etc. (Martin & Eklund, 2001). Subgraph isomorphism detection and the organization of knowledge into various kinds of specialization or partOf hierarchies according to the results of graph-matching operators permit more efficient information retrieval or logic deduction than classic logic-based methods (Ellis, 1995) (Messmer & Bunke, 2000). Presenting knowledge in a high level way and organizing it into hierarchies for query results is also more efficient if that knowledge is already stored that way instead of using a low-level model such as KIF or even RDF+OWL. This is why the data model of WebKB-2 has been extended whenever the proposed high-level notations have been extended. More generally, for a programmer as well as an end-user, managing knowledge via high-level models or notations is much easier than with low-level ones. The code of WebKB-2 has been designed to be generic enough for various back-ends (e.g., DBMSs, remote knowledge servers, main-memory inference engines) to be used instead of the currently used DBMS.
- However, my past research works and WebKB-2 are only a basis for the above mentioned goals. Additional **graph-matching operators** should be defined. This is of particular interest to KEWI (Corby & Faron-Zucker, 2007) (Corby et al., 2007). The model and proposed notations should also include **higher level constructs** (for example, a construct to represent natural language expressions such as "3 by 3" as in "bringing objects 3 by 3" and other collection related constructs which most people would find very difficult to represent correctly in logic and which most inference engines would be unable to exploit) and these constructs should be taken into account by the graph-matching operators. No **external inference engine** has so far actually been exploited by WebKB-2. With respect to the above cited goals and some subgoals of the "knowledge sharing and exchange" theme, it would be very interesting to exploit (i) classic description-logic based inference engines via de-facto standards such as the Protégé-OWL reasoner API, and (ii) first-order logic reasoners such as Vampire which is currently much re-used for reasoning with OWL-Full or SWRL (a Semantic Web Rule Language Combining OWL and RuleML). I designed For-Links (FL) and Frame Conceptual Graphs (FCG) to provide the most concise, normalized, structured and expressive notational versions for, respectively, description-logic based textual notations and Nth-order logic textual notations (Martin, 2002). These notations are interesting basis for the **generic knowledge entering/display/exchange notations** that KEWI aims to design.

Fully and correctly importing or exporting between these notations and other languages such as such as KIF, CL and W3C languages such as RDF+OWL and N3 is a difficult but important research work since it means translating them from/to languages with different structures (lower-level structures or non-graph based structures) and different expressiveness. Yao & Etkorna (2006), or Dieng-Kuntz & Corby (2005) of the INRIA Edelweiss team with which the KEWI team regularly collaborate, have already proposed methods to convert knowledge between the RDF model and its related subset within the Conceptual Graph model. Taking more expressiveness into account implies making more ontological choices and exploiting as well as complementing the ontology on which the source and destination knowledge representations (KRs) are based. Thus, this research involves ontology translation (Corcho, 2005) (Euzenat & Shvaiko, 2007) and hence overlaps with the one on ontology integration. This research will be done in collaboration with the EXMO team (led by Euzenat and including Corcho) and will be guided by **methodological and normalization guidelines and protocols**. To come up with them and, more generally, to guide knowledge entering, the current guidelines and protocols of WebKB-2 will be refined and extended. For example:

- A better detection of inconsistencies and partial redundancies is needed, and different kinds of inconsistencies or partial redundancies should be not only treated differently but in accordance to each user's preferences.
- My complementary system of "category cloning" (Martin, 1996) should be refined and integrated.
- It is also necessary to integrate and represent concepts and ideas from ontology engineering methodologies such as Methontology, DOGMA, Diligent and HCOME.
- Privacy issues should be taken into account. On this issue too a collaboration with EXMO can be expected. The investigation will start with protocols permitting a user to specify rules to hide some of her information to some/all other users until they send her certain specified kinds of information; this will be important for the use of KB servers for auction sites or "social spaces". For social networking, Weitzner (2007) also stresses the need of letting users associate usage/inference policy restrictions to their data (e.g., "do not remove the usage or inference restrictions associated to these data", "do not use these data for employment decisions" and "do not infer religious affiliation from these data"). With the approach I propose, such rules could be specified in a precise and flexible manner (by using or extending the ontology) and some usages could be automatically checked (e.g., assertions about the religious affiliation of someone based on data to which a "do not infer religious affiliation" restriction has been associated).
- Additional mechanisms need to be found to detect or guide the resolution of independent extensions to a KB when these extensions could (and hence should) be merged into more precise and normalized KRs by using different conceptual categories (this problem is an ubiquitous one which requires going far beyond current techniques for ontology merging and evolution).

Formalized English (FE) is currently only a syntactic variant of FCG and hence does not lead to translation research problems. The downside is that this FE is not as easy to read or write as other formal or informal **controlled languages** which are often proposed to ease knowledge representation. A first necessary extension is, while keeping FE formal, to allow it to be more English-like by extending its grammar and exploiting the ontology. Then, via a set of options, each user should be allowed to see or use **more and more informal extensions to FE**. Ultimately, the use of simple but common English sentences should be allowed for knowledge entering and querying, with their translations into a formal version of FE being shown to the user for her to check the interpretation being made of her sentence or select between different interpretations. This option-based approach will offer the advantages of all "controlled languages", formal and informal. A last option will allow the parsing of several sentences without the user having to check the interpretations; in this case, a semi-formal version of each sentence will be stored. This means that the data model and the graph-matching operators will have to be updated to accept more informal elements. Words (instead of category identifiers) are already accepted by WebKB-2 but, for scalability and efficiency purposes, the data model needs to be made more generic (e.g., my recent investigations led to the conclusion that words and category identifiers should not be stored into different data structures and should be organized into a *same* specialization hierarchy - unlike in WordNet for example). This research overlaps with the one on knowledge extraction from natural language described in Section 5.

4. Knowledge presentation

The results of my research ease the collaborative or personal representation and evaluation of formal and informal information (e.g., argumentation structures and information about users, including their contributions) in a precise and uniform way. The represented information should then be usable for knowledge querying, filtering and presentation purposes. WebKB-2 already proposes many presentation options and permits the specification of filtering constraints using FL or FCG. For example, during browsing or in the results of queries, certain categories or statements created by authors having certain characteristics are not displayed or are displayed in small fonts. However, the filtering specifications must be manually made, and no set of commands nor conceptual categories are proposed to permit the user to specify a presentation. Hence, it is interesting for ScallIntegr to complement and refine my knowledge-presentation related research by investigating (i) a presentation specification language (ontology + commands) and (ii) automatic methods to acquire presentation specifications. This will be useful for the research on "the adaptation of knowledge based on users' profiles" of the KEWI team, or the research on "ergonomic interactive systems" and "models of written and argumentative communication" of the IC3 team. Unlike Fresnel (Pietriga et al., 2006) the presentation specification language will only deal with simple presentation details (such "what kind of information to put in bold characters"), not with graph layout details, but will be integrated in the languages of commands of WebKB-2 and hence the commands will be easy to combine (for example via "pipes" as in Unix) and associate to hyperlinks.

Even for the Semantic Learning Web (Stutt & Motta, 2004) and Educational Semantic Web (Devedzic, 2004) the exploited or acquired user models are very coarse (e.g., grade level and mastery level in certain courses) and most often predefined. The use of a large (virtual) shared KB permits to store what people have seen or been tested on (and when) in a precise and uniform way.

In addition to presenting information in personalised ways, presenting it in an organized way is important too. Indeed, (i) this permit people to see and find what may be of interest to them in a more efficient way, (ii) this may avoid the need to filter information (even manually specified filtering may cause interesting information to be missed, and it is dangerous to let an automatic mechanism guess what is interesting or not for a user), and (iii) this avoids the need to give a list of "related information that might also be of interest". Knowledge normalisation guidelines and protocols help achieving or keeping a good organization but additional mechanisms are needed to structure query results. For example, a scalable way to compare objects should be found and developed; an idea to start with is illustrated in Table 2 and presented in (Martin et al., 2005). Long lists of query results should also be structured into conceptual hierarchies.

For this last goal, one approach consists in structuring an initially non-organized long list of statements satisfying a query according to specialization relations that can be calculated or that have been set between the statements. The list thus becomes a list of statement hierarchies the nodes of which a user can navigate, contract or expand. A complementary approach that would require more research and experiments would be to structure the statements of the list according to various topics to which these statements are related. Examples of topics that can be used to group statements about a certain person are: physical characteristics, administrative details, possessions, work related activities, etc. I believe that such groups (and their sub-hierarchies) can be extracted based on either the type of the first relation of each statement, or if this type is too common, the type of the destination concept node of the first relation. It may turn out that be the second relations and their destinations, or other complementary information, will also have to be exploited.

Table 2. Generation or exploitation of scalable semantically-structured comparison tables.

This example first presents two normalized FCG statements about two ontology tools (Ontolingua and WebKB-2), then a command to compare these tools on their support of KR notations, and finally its result. This result is a table where

- rows are features organized into a specialization hierarchy,
- each column corresponds to a tool (possibly organized into a specialization hierarchy),
- each cell uses a mark to indicate whether the corresponding tool has the corresponding feature ("+"), nearly has it (e.g., "~2009"), does not have it ("-"), is not known by the users of the KB to have it ("."), etc.

Such tables can also be used for knowledge entering purposes. As a service to the Standard Upper Ontology community and the Conceptual Graph communities, I used such tables (with the above cited marks and more precise ones) to assert and compare features of KR notations (currently, 10 families of notations according to 51 features) as well as features of KR tools (currently 6 tools according to 160 features). Such tables - which I named "spec-tables" - are scalable because the features (and sometimes the compared objects too) are organized into a specialization hierarchy and hence new features or tools can be added without changing the overall organization of the table. This is not possible when a cell of the table can contain one or several feature names and values, as is for example the case in Fact Guru, one of the rare KBMSs able to generate comparison tables.

Please note that in the example below the feature within parenthesis refers to a type of features that has not yet been named (i.e., no category has been entered to represent this particular type); the description of this type is automatically generated to represent the fact that the two tools share that types of features even though they do not share the two sub-features.

An interesting extension to investigate regarding these tables would be the use of partOf relations (in addition to specialization relations) to organize the features.

Finally, please note that in the following representations, for readability purposes, the source or creator of each category is not indicated: no category identifier includes the identifier of its creator or source, as a prefix or suffix.

```
[any Ontolingua, output_language: a KIF, support of: a lexical_search,
      part: {a HTML_based_interface, no graph_visualization_interface}};
//the first line of this FCG can be read: any (instance of the KB server) Ontolingua has
// for output_language (an instance of the notation) KIF and is a support of lexical_searches
```

```
[any WebKB-2, output_language: {a FCG, a RDF},
      part: (a user_interface, part: {an API, a HTML_based_interface,
      a CGI_accessible_command_interface,
      no graph_visualization_interface}),
      support of: (a specialization_structural_retrieval,
      kind: {complete_inferencing, consistent_inferencing},
      input: (a query, expressiveness: PCEF_logic),
      object: (several_statement, expressiveness: PCEF_logic)
      )]; // "PCEF": positive conjunctive existential formula
```

```
compare pm#WebKB-2 km#Ontolingua on
      (support of: a IR_task, output_language: a km#KR_notation,
      part: a user_interface), maxdepth 5
```

output_language:	WebKB-2	Ontolingua
<i>KR_notation</i>	+	+
XML-based notation	+	-
RDF	+	-
(expressiveness: FOL)	+	+
FCG	+	-
KIF	.	+
part:		
<i>user_interface</i>	+	+
HTML_based_interface	+	+
CGI_accessible_command_interface	+	.
OKBC_interface	.	.
API	+	.
graph_visualization_interface	-	-

5. Semi-automatic ontology integration and knowledge extraction from English sentences

As previously noted the originality of ScalIntegr's research on (semi-)automatic ontology integration and knowledge extraction from English sentences (KEE) is that both tasks will be conducted jointly and in the context of ScalIntegr's other research tasks (i.e., knowledge integration into a large shared KB, knowledge normalisation guidelines and protocols, replication of knowledge between knowledge bases, progressive extension of Formalized English and of the data model of WebKB-2, and support for the collaboratively-built semi-formal states of the art). Conducting these two tasks "jointly" means that they will share similar techniques, that for ontology integration some KEE procedures will be applied on the names, annotations or other informal information associated to conceptual categories, and that the result of the use of these techniques will increase the shared KB and hence will give more background knowledge for these techniques to work. To that end, the integration of ontologies such as FrameNet (an ontology of English verbs), Extended WordNet (WordNet with pre-parsed annotations) and DBpedia (a base of structured information extracted from Wikipedia) will be prime targets. The result will be easily accessible by researchers or applications like the current Multi-Source Ontology of WebKB-2 is today. Finally, as part of the initialization of an ontology for a collaboratively-built semi-formal state of the art in Ontology Engineering, the main techniques of ontology merging and natural language understanding will be represented. Given that (i) all the tasks of ScalIntegr are inter-related, (ii) ScalIntegr starts from my existing work on a shared KB approach and its partial integration of my work for the "semi-independent approach" (WebKB-2 includes WebKB-1), the research presented in this section is an important part of this long term project.

This research fits the "knowledge extraction and integration" and "Models of knowledge and texts" research themes of, respectively, KEWI (I3S laboratory) and IC3 (IRIT laboratory). Furthermore, I3S includes the team "RL" (for Ressources Linguistiques) while Natural Language Understanding (NLU) is a transversal research theme of the IRIT laboratory.

The investigation of semi-automatic ontology integration will start by re-using and extending the techniques and softwares of EXMO for converting knowledge between various languages, aligning ontologies, reconciling discrepancies, performing terminological analysis, etc. Examples of resulting softwares are the Transmorpher 1.0, the API for ontology and the OWL-Lite Alignment tool (Djoufak-Kengue et al., 2007).

Regarding KEE, ScalIntegr will re-use NLU parsers from the INRIA-CNRS UMR7503 (LORIA) but will adapt them to exploit the content of the shared KB and generate knowledge representations that are more precise and normalised than they currently are. For example, the NLU parser should be able to translate "governments should enforce animal rights" into the normalized FE sentence 'the enforcement_of_animal_rights_by_governments should happen' the following FE definition for the concept type "enforcement_of_animal_rights_by_governments": 'an enforcement with agent some government and with object some animal_rights'. Then, if the user accepts, or if the translation is automatic, the two knowledge representations may be inserted into the KB. Whether actually inserting categories with names such as "enforcement_of_animal_rights_by_governments" is a good idea or not is another research issue; this depends on various factors such as (i) the existence or not of many statements about such a category, (ii) the domain of the KB, and (iii) the possibility or not to later regenerate the definition of the category from its name. Although many sentences will not be fully automatically understandable, normalizing and indexing many of their components (and especially the processes they refer to) will go a long way toward permitting the normalization, interconnection, classification or comparison of statements and other objects. Similarly, since the most important knowledge to extract are basic relations between categories such as subtypeOf, instanceOf, partOf and substanceOf relations, this will be the main goal of ScalIntegr's KEE related investigation. LORIA has many NLU related research projects, also intends to exploit FrameNet, and thus would be a strong partner for this research.

In the KEWI team, bioinformatics related applications are likely to arise. In that context, it is necessary to re-use specialized bioinformatics KEE tools. The Natural Language Processing Group of the Medical Informatics Division of the University Hospital of Geneva have agreed to cooperate with me on extensions

of WebKB-2 for bioinformatics related applications and have allowed me to have the sources of GALEN (Trombert-Paviot et al., 2000) which is (one of) the most advanced of medical text parsers.

6. Main test applications: collaboratively-built semi-formal repositories (states of the art, corporate memories, ...)

Overview. My research report and this document have introduced complementary techniques which, when combined, offer a scalable solution to the collaborative building of a "virtual" well-organized KB based on one or several KB servers on the Web or within an intranet or peer-to-peer network (semantic grid). Such KB can be formal (like CYC for a general KB or OpenGALEN for a medical-oriented KB) or semi-formal. It can be used to progressively organize the large amount of unstructured or semi-independently structured information created by communities via their mailing lists, forums, publications or ontologies. With the tools currently used by communities, the organization of the information is very coarse-grained: most often, whole documents or ontologies are indexed (their statements are not organized into one semantic network). Information on what people know, prefer, can provide or have provided is rarely stored. A main goal of KEWI is to address these organization restrictions (and thereby collaboration restrictions) especially for "organizational memories", the "social and collaborative Web" and (e-)learning (Dehors et al., 2006). Among other tools developed by KEWI to achieve that goal is SweetWiki (Buffa et al., 2007) a user-friendly semantic wiki marrying ontologies and folksonomies. This tool is complementary to WebKB-2 which can also be seen as a semantic wiki but only based on a tight semantic network, not folksonomies. Both approaches can and should be integrated. Part of my role in the KEWI team would be to federate, complement and apply its research in the context of projects, especially industrial projects or European projects - and also find or organize such projects. Projects involving corporate memories, bioinformatics applications, (e-)learning applications or tourism-related applications will be prime targets. I have some concrete experience of using WebKB-2 for the last two kinds of applications (Martin, 2005, 2006). Other interesting domains of application that will particularly test the distribution and cooperation approaches described in the previous sections and in my research report will be e-government and e-sciences. In addition to those applications ScalIntegr will permit people to collaboratively build states of the art and complements to Wikipedia in ways that better help researchers, lecturers, students, engineers and decision makers.

Semantically structured complements to Wikipedia for technical or original materials and hotly debated issues. The goals and approach of Wikipedia restrict it to be poorly structured, of a general "encyclopaedic nature" and "consensual". These restrictions are not lessened by the existence of sister-projects (Wiktionary, Wikiquote, Wikibooks, etc.) nor the potential use of [Semantic MediaWiki](#) in the future. The approach I propose permits to go beyond these restrictions and, more generally, beyond certain restrictions of the many on-going research works aiming to improve Wikipedia (e.g., they do not fully address the problems related to the existence of a committee or to right of removal by any user, nor do they provide ways to normalize semi-formal semantic networks and keep them well-organized). Hence, ScalIntegr's approach could be used as a complement for Wikipedia. Using fine-grained argumentation structures, a first step is to represent the debates of many Wikipedia pages marked with having a debated content and, at the bottom these pages, will invite people to explore and complement the argumentation structures in WebKB-2. The intellectual challenge and opportunity for recognition may attract a lot of people. This challenge may be viewed as one of the "games masquerading core tasks of weaving the Semantic Web behind on-line, multi-player game scenarios" such as OntoGame (Siorpaes & Hepp, 2007). I have begun the representation of some of these debates (Martin, 2006b).

Semi-formal states of the art. The beginning of Section 2 of my research report lists reasons why storing formal, semi-formal or informal information into personal documents or KBs was a restricting and sub-optimal way of sharing information. Here are some complementary (although related) reasons why the current system of publishing research outputs via informal research articles is sub-optimal for knowledge sharing compared to the collaboratively building of a semi-formal state of the art.

- Writing such articles implies repeating in different ways information that have been described elsewhere, for example to summarize ideas, appear original or avoid copyright issues. Refinements of previously published ideas cannot be simply associated to those ideas, for example via specialization or correction relations; instead, a whole new article has to be written. Given the increasing number of newly published articles, it is difficult and time-consuming for a person or a software to find, relate, compare and evaluate the various presented ideas, techniques or tools. The problems are compounded by the limited memory of people and the fact that softwares do not understand the content of documents.
- Writing such articles does not permit to organize concepts or ideas into structured forms (as for example the argumentation structure illustrated in Table 3 of my research report) thus leading to repetitions as well as less rigorous and more superficial argumentations (reading or writing a linear form of a deep argumentation structure would be difficult).
- Writing such articles involves making choices and compromises about what to describe and how (level of detail, order, etc.) according to the expected knowledge and interests of the readers, and assumed expectations of the reviewers for the readers. These expectations are often incorrect. In any case, most readers will have to read information they already know and will want further details that other readers will not want.
- The articles that are most likely to be published in selective journals or conferences may well be those that only include well-known uncontroversial ideas (Sowa, 2007) and, more generally, easy-to-understand not-too-technical ideas. Even if the reviewer of an article is as knowledgeable as its author about the presented ideas, and hence can see past a one-sided presentation of those ideas, the classic review and publication system does not permit reviewers to associate a valuation, argument or counter-argument to any particular idea of the article nor does it permit the author to give any feedback to correct mis-interpretations.
- The content of research articles cannot be automatically re-used to generate documents for other purposes, e.g., e-learning.

As noted in the introduction, the techniques and research directions the project ScalIntegr will permit to solve these problems and hence will provide a way to fulfill the dream of Hillis (2004) and Buckingham et al. (2007). The actual fulfillment of this dream will then be a social issue, not a technical issue. It is clear that the current WebKB-2 and its current notations are too difficult to use for most researchers and lecturers to participate in representing states of the art but the proposed extensions and the use of a graphical interface (or a structured document interface for KRs) will solve those problems.

It should be reminded that the approach of the project ScalIntegr allows semi-formal information and hence permits an incremental organization of the information. Thus, from a technical viewpoint, a cooperatively-built well organized state of the art is achievable in the short to medium term. By contrast, the [Halo project](#) (Friedland et al., 2004) is a much more ambitious and long term project since its goal is to create a "Digital Aristotle" capable of teaching much of the world's scientific knowledge and also using it to solve classic exercises, thus requiring a fully formal KB supporting complex problem solving mechanisms. Such formal KBs permit to support problem solving but they are not meant to be directly read or browsed, and designing them is difficult even for teams of trained knowledge engineers, e.g., the six-month pilot phase of Project Halo was restricted to only 70 pages of a chemistry book and had encouraging but far-from-ideal results.

From a content-oriented viewpoint, ScalIntegr's first goal regarding states of the art is to continue the initial ontology of Knowledge Engineering that I have begun (Martin, 2007) - including structured discussions about debates in this field, such as those I currently create and invite the Standard Upper Ontology (SUO) community to contribute to in the SUO wiki that I administrate - until it is a sufficient guide for scalable additions, that is, when it is easy for researchers to know where to insert new categories or statements for the KB to remain well organized. When this time comes, researchers in Knowledge Engineering will be invited to complete this ontology.

As shown in (Martin & Eboueya, 2008), given its structure, this state of the art will be the ultimate (e-)learning resource in this field. Even for the individual courses that I represented, the students of these

courses recognized the help that the semantic network provided them in relating and comparing information scattered in the different slides and other lecture materials of these courses. Furthermore, the possibility for the students to complement the semantic network permit precise feedback by the students and provide their teachers a way to evaluate their knowledge and analytic skills, directly or via the cooperative evaluation method that this project should refine (Section 2.3 of my research report) (Martin et al., 2006). More generally, this approach is in-line with the constructivist and argumentation theories and can be seen as a particular implementation and support of the "critical thinking" theories and Brandom's model of discursive practice (Brandom, 1998).

Semi-formal state of the art in Ontology Engineering. No semi-formal state of the art of Knowledge Engineering (KE) currently exists. The need for it is however clear in the KE community. An example of this need is the interest that [Michael Denny's "Ontology editor survey"](#) (Denny, 2004) attracted despite the fact that its very coarse level of details made it quite unusable for its main goal: tool comparison. Another example is the [Semantic Web Topics Ontology](#) of ISWC 2006. However, this ontology is very poorly structured (it is by no means helping to find a right place or way to insert information) and is only aimed to support the indexation of documents. The initial ontology of KE that I have begun is not an ontology of topics (the classification of topics cannot be normalised, it is task-dependant), it is an ontology of KE tasks, data structures, notations, tools, and their properties (Martin & Eboueya, 2007).

Semi-formal state of the art in Software Engineering. Creating an ontology of Ontology Engineering leads to include many Software Engineering related categories. One of the goals of the KEWI and RAINBOW teams is to represent knowledge about Software Engineering. Hence, the team members of KEWI and I will extend this ontology.

7. References

Brandom, R. (1998). Action, Norms, and Practical Reasoning. *Noûs*, Volume 32, Supplement 2, pp. 127-139.

Buckingham-Shum S., Motta E. & Domingue J. (1999). *Representing Scholarly Claims in Internet Digital Libraries: A Knowledge Modelling Approach*. ECDL 1999 (pp. 423-442), 3rd European Conf. Research and Advanced Technology for Digital Libraries, Paris, France, September 1999.

Buffa M., Gandon F. & Ereteo G. (2007). *A wiki on the semantic web*. Chapter of a book titled "Emerging technologies for semantic web environments: techniques, methods and applications" and edited by Jörg Rech, Björn Decker and Eric Ras, Fraunhofer of the Institute for Experimental Software Engineering (IESE), Germany, July 2007.

Corby O. & Faron-Zucker C. (2007). *Implementation of SPARQL Query Language based on Graph Homomorphism*. Proceedings of ICCS 2007, 15th International Conference on Conceptual Structures, Sheffield, UK.

Corby O., Dieng-Kuntz R., Faron-Zucker C. & Gandon F. (2007). *Searching the Semantic Web: Approximate Query Processing based on Ontologies*. IEEE Intelligent Systems Journal, Vol. 21, No. 1, January/February 2006.

Corcho O. (2005). *A layered declarative approach to ontology translation with knowledge preservation*. Frontiers in Artificial Intelligence and its Applications. Dissertations in Artificial Intelligence. IOS Press. January 2005.

Berners-Lee T., Hendler J. & Lassila O. (2001). *The Semantic Web*. Scientific American, May 17th, 2001.

De Roure D., Jennings N.R. & Shadbolt N.R. (2005). *The Semantic Grid: Past, Present and Future*. Proceedings of the IEEE, vol. 93, no.3, pp. 669-681, March 2005.

Dehors S., Faron-Zucker C. & Dieng-Kuntz R. (2006). *Reusing Learning Resources based on Semantic Web Technologies*. Proceedings of ICALT 2006, 6th IEEE International Conference on Advanced Learning Technologies, Kerkrade, The Netherlands, July 2006.

Denny M. (2004). *Ontology Tools Survey, Revisited*. Web document created on July 14th 2004. <http://www.xml.com/pub/a/2004/07/14/onto.html>

Devedzic, V. (2004). Education and the Semantic Web. *International Journal of Artificial Intelligence in Education*, 14, pp. 39-65.

Dieng-Kuntz R. & Corby O. (2005). *Conceptual Graphs for Semantic Web Applications*. "Conceptual Structures: Common Semantics for Sharing Knowledge", LNCS 3596/2005, pp. 19-50.

Djoufak-Kengue J.-F., Euzenat J. & Valtchev P. (2007). *OLA in the OAEI 2007 evaluation contest*. Proceedings of ISWC 2007 workshop on ontology matching, pp. 117-184, Busan, Korea.

Ellis G. (1995). *Compiling Conceptual Graphs*. IEEE Transactions on Knowledge and Data Engineering, February 1995, Vol. 7, No. 1, pp. 68-81.

Euzenat J. & Shvaiko P. (2007). *Ontology Matching*. Springer-Verlag, Berlin Heidelberg (DE), 2007, 341 pages.

Friedland N.S., Allen P., Mathews G., Whitbrock M., Baxter D., Curtis J., Shepard B., Miraglia P., Angele J., Staab S., Moench E., Opperman H., Wenke D., Israel D., Chaudhri V., Porter B., Barker K., Fan J., Chaw S.Y., Yeh P., Tecuci D. & Clark P. (2004). *Project Halo: Towards a Digital Aristotle*. AI Magazine, 25(4), 2004, pp. 29-48.

Hillis W.D. (2004). *"Aristotle" (The Knowledge Web)*. Edge Foundation, Inc., No 138, May 6, 2004.

Kalfoglou Y., Alani H., Schorlemmer M. & Walton C. (2004). *On the Emergent Semantic Web and Overlooked Issues*. Proceedings of ISWC'04, 3rd International Semantic Web Conference, LNCS 3298, Hiroshima, Japan, pp. 576-, November 2004.

[Martin Ph. \(1993\)](#). *A KADS refinement for Explanatory Knowledge Extraction and Modelling*. Proceedings of AI 1993, 6th Australian Joint Conference on Artificial Intelligence (edited by "World Scientific, Singapore"), Melbourne, Australia, November 16-19, 1993.

[Martin Ph. \(1996\)](#). *Exploitation de graphes conceptuels et de documents structurés et hypertextes pour l'acquisition de connaissances et la recherche d'informations*. Ph.D. thesis (378 pages), University of Nice - Sophia Antipolis, France, October 14, 1996.

[Martin Ph. \(2000\)](#). *Conventions and Notations for Knowledge Representation and Retrieval*. Proceedings of ICCS 2000, 8th International Conference on Conceptual Structures (Springer, LNAI 1867, pp. 41-54; electronically published on 1/1/2007), Darmstadt, Germany, August 14-18, 2000.

[Martin Ph. & Eklund P. \(2001\)](#). *Large-scale cooperatively-built heterogeneous KBs*. Proceedings of ICCS 2001, 9th International Conference on Conceptual Structures (Springer, LNAI 2120, pp. 231-244), Stanford University, California, USA, July 30 to August 3, 2001.

Martin Ph. (2002). *Knowledge representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English*. Proceedings of ICCS 2002, 10th International Conference on Conceptual Structures (Springer, LNAI 2393, pp. 77-91), Borovets, Bulgaria, July 15-19, 2002.

Martin Ph. (2003). *Correction and Extension of WordNet 1.7*. Proceedings of ICCS 2003, 11th International Conference on Conceptual Structures (Springer, LNAI 2746, pp. 160-173), Dresden, Germany, July 21-25, 2003.

Martin Ph., Blumenstein M. & Deer P. (2005). *Toward cooperatively-built knowledge repositories*. Proceedings of ICCS 2005, 13th International Conference on Conceptual Structures (Springer, LNAI 3596, pp. 411-424), Kassel, Germany, July 18-22, 2005.

Martin Ph. (2005). *Services on the Sunshine Coast*. <http://www.webkb.org/kb/SC/>

Martin Ph. (2006). *Documents related to my Griffith E-Learning Fellowship for Semester 2, 2006*. <http://www.webkb.org/doc/papers/GEL06/>

Martin Ph. (2006b). *Structured discussions & Semantic classification of some resources*. <http://www.webkb.org/kb/it/>

Martin Ph., Eboueya M., Blumenstein M. & Deer P. (2006). *A Network of Semantically Structured Wikipedia to Bind Information*. Proceedings of E-learn 2006 (pp. 1684-1702), AACE Conference on E-learning in Corporate, Government, Healthcare, & Higher Education, Honolulu, Hawaii, October 13-17, 2006.

Martin Ph. & Eboueya M. (2007). *Sharing and Comparing Information about Knowledge Engineering*. WSEAS Transactions on Information Science and Applications, Issue 5, Volume 4 (pp. 1089-1096; ISSN: 1790-0832), May 2007.

Martin Ph. & Eboueya M. (2008). *For the ultimate accessibility and re-usability*. Book chapter to be published early 2008 in the Handbook of Research on Learning Design and Learning Objects: Issues, Applications and Technologies.

Messmer B.T. & Bunke H. (2000). *Efficient Subgraph Isomorphism Detection: A Decomposition Approach*. IEEE Transactions on Knowledge and Data Engineering, March 2000, Volume 12, Issue 2, pp. 307 - 323.

Noy N.F., Chugh A., Liu W. & Musen M. A. (2006). *A framework for ontology evolution in collaborative environments*. Proceedings of ISWC 2006, 5th International Semantic Web Conference, Athens, GA, 2006.

Patel-Schneider P.F. (2005). *A Revised Architecture for Semantic Web Reasoning*. Proceedings of Third Workshop on Principles and Practice of Semantic Web Reasoning (PPSWR'05), Dagstuhl, Germany (11th - 16th September 2005), LNCS 3703, pp. 32-36, September 2005.

Pietriga E., Bizer C., Karger D., Lee R. (2006). *Fresnel: A Browser-Independent Presentation Vocabulary for RDF*. Proceedings of ISWC 2006, 5th International Semantic Web Conference, LNCS 4273, pages 158-171, November 2006, Athens, GA, USA.

Shadbolt N., Berners-Lee T. & Hall W. (2006). *The Semantic Web Revisited*. IEEE Intelligent Systems, 21(3) pp. 96-101, May/June 2006.

Rousset M-C. (2004). *Small Can Be Beautiful in the Semantic Web*. Proceedings of International Semantic Web Conference (ISWC 2004), pp. 6-16.

Siorpaes K. & Hepp M. (2007). *OntoGame: Towards Overcoming the Incentive Bottleneck in Ontology Building*. Proceedings of the 3rd International IFIP Workshop On Semantic Web & Web Semantics (SWWS 2007), LNCS 4806, pp. 1222-1232, November 29-30, 2007.

Stutt A. & Motta E. (2004). Semantic Learning Webs. *Journal of Interactive Media in Education*, Special Issue on the Educational Semantic Web, 10.

Trombert-Paviot B., Rodrigues J.M., Rogers J.E., Baud R., van der Haring E., Rassinoux A.M., Abrial V., Clavel L. & Idir H. (2000). GALEN: a third generation terminology tool to support a multipurpose national coding system for surgical procedures. *International Journal of Medical Informatics*, Vol. 58-59 (pp. 71-85), September 1st 2000.

Vrandecic D., Pinto H.S., Sure Y. & Tempich C. (2005). *The DILIGENT Knowledge Processes*. *Journal of Knowledge Management*, 9(5), pp. 85-96, October 2005.

Weitzner D. (2007). *Reciprocal Privacy (ReP) for the Social Web*. Web document created on December 12th 2007. <http://dig.csail.mit.edu/2007/12/rep.html>

Yao H. & Etkorna L. (2006). *Automated conversion between different knowledge representation formats*. *Knowledge-Based Systems*, Volume 19, Issue 6, October 2006, pp. 404-412.