Toward the ultimate precision, modularity and structure for the ultimate accessibility, re-usability learning and evaluation support

Dr Philippe Martin[1], Dr Michel Eboueya[2]
1: School of I.C.T. - Griffith University, Australia; *2*: L3I, University of La Rochelle, France
This work was supported by Griffith University and a grant from the Region Poitou-Charentes (France)

Abstract. We first argue that current approaches for sharing and retrieving learning objects or any other kinds of information are not efficient or scalable, essentially because almost all of these approaches are based on the manual or automatic indexation or merge of *independently created* formal or informal resources. We then show that tightly interconnected *collaboratively updated* formal or semi-formal large knowledge bases (semantic networks) can, should, and probably will, be used as a shared medium for researching, publishing, teaching, learning, evaluating or collaborating, and thus ease or complement traditional methods such as face-to-face teaching and document publishing. To test and support our claims we have implemented our ideas into a knowledge server named WebKB-2 and begun representing our own research domain and several courses at our university. The underlying techniques could be applied to a semantic/learning grid or peer-to-peer network.

Keywords: knowledge representation/normalization/sharing/retrieval/learning/evaluation, design principles for learning objects, learning object re-usability/interoperability/repositories

Table Of Contents

1. Introduction

The smaller and less contextual the "learning objects (LOs) available for re-use" are, and the more precisely indexed or inter-connected via metadata they are, the more easily they can be semi-automatically retrieved and combined to create "LOs to teach with" that are adapted to particular course objectives or kinds of users, and thus create contextual LOs (Downes, 2001; Hodgins, 2006). Although this general idea is well advocated in the LO community, its ultimate conclusion - the idea that we advocate - is hardly attempted or even written about: each "re-usable LO", which from now on is simply referred to as an "object", should either be one formal term (a category identifier) or an "un-decomposable statement" (typically, one semantic relation between two other objects, with some information about the context of this relation, such as its creator and temporal, spatial or modal constraints on its validity, all of which preferably being expressed in a formal way, that is, with a knowledge representation language). Furthermore, each object should be connected to all other semantically related objects by semantic relations. In other words, there should be no difference between data and metadata, and there should be only one virtual well-organized knowledge base (KB) that all object providers can complement by inserting their objects "at the right place", or more generally, in a "normalized way" that permits the KB to stay well organised and hence to be searched and updated in an efficient or scalable way. A virtual KB does not imply only one actual KB, it simply means that all potential redundancies and inconsistencies detected by people or inference engines should be removed. As explained in Section 3.1, this also does not imply that knowledge providers have to agree with each other.

Nowadays, there is no such virtual KB, and LOs repositories are not even KBs, they are databases for informal documents containing many more than one un-decomposable statement. Furthermore, current LO related standards (e.g., AICC, SCORM, ISM, IEEE WG12) and projects (e.g., CANDLE, GEODE, MERLOT, VLORN) essentially focus on associating *simple metadata* to whole documents or big parts of them (e.g., author, owner, terms of distribution, presentation format, and pedagogical attributes such as teaching or interaction style, grade level, mastery level and prerequisites). Such superficial indices do not support the answering of queries 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?". Answering such queries requires presenting and allowing the browsing of the KB as 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.

LOs have special purposes but no special content: all advanced information sharing or retrieval techniques can be directed applied to LOs. On the Web, this means using *Semantic Web* related techniques (Shadbolt, Berners-Lee, & Hall, 2006). However, almost all them are about supporting the manual/automatic indexation of whole formal/informal documents or merging the content of independently created formal documents. Document-based techniques permit to exploit legacy data but their efficiency or scalability for organising, sharing and searching increasingly large amounts of information is limited. Hence, these techniques should ideally be used only as a complement to the building of a global virtual KB, not as sole techniques for exploiting information. This is the theme of the next section. Then, we show how such a virtual KB - on the Web or within the semantic/learning grid of a community - can and ultimately will be collaboratively built and hence used as a shared medium for researching, publishing, teaching, learning or collaborating since these tasks are based on information retrieval/comparison/sharing subtasks.

2. Background: Current Information Retrieval/Sharing Approaches Are Not Scalable

Definitions. In this article, a "formal term" is a symbol (character string, icon, sound, etc.) whose meaning (i.e., the referred concept/relation type/individual) has been made explicit, a "statement" is a small set of symbols connected by relations, an "informal statement" is a statement without formal terms (e.g., a sentence in English), a "formal statement" is a statement with only formal terms, a "semi-formal statement" is a statement with formal relations and may be formal terms for concepts or individuals, an "object" (or re-usable LO) is either a term or a statement, an "ontology" is a set of formal objects (e.g., a small flat list or a full KB), a "resource" is a stand-alone collection of several statements (e.g., an ontology, a database, a document, a section or a paragraph), and "metadata" is a set of one or several numerical values or other objects used for relating or indexing one or more statements, typically those of a resource. Some metadata related to some resource or created by some person(s) can also be considered as a resource. Scalability means keeping precision-oriented information retrieval/comparison/sharing efficient even when the number of statements written by all the information providers grows large. Section 7 defines (scalable) knowledge sharing, normalization, comparison and retrieval.

2.1. Approaches Based on the Indexation of Resources Are Not Scalable

The more statements a resource contains, and the more resources there are, the more these resources contain similar and/or complementary pieces of information, and hence the less the metadata for each resource can be useful: queries will return lists of resources that are partially redundant or complementary with each other and that need to be manually searched, compared or aggregated by each user. Furthermore, the more statements a resource contains, the more its metadata have to be information selective, and hence the less such metadata are representative of the contained pieces of information and the more the indexation methods and usefulness are task/user/domain dependent.

Finally, the more statements some resources contain, and the less formal the statements are (or the more "contextual" they are), the less any similarity measure between these resources can have any intuitive or semantic meaning, and the less these resources can meaningfully be related by rhetorical or argumentation relations such as "arguments", "proves" or "specializes". For example, the statement "some animal sits above some artefact" is a generalization (i.e., logical implication) of both "Tom (a cat) sits on a blue mat" and "any animal sits above some artefact" because all the objects and quantifiers of the first statement are identical or generalize those of the second and third statements (such relations can be automatically inferred if the statements are formal or semi-formal). However, such relations rarely hold between two collections of statements, and especially between any two documents. Statistical similarity measures between documents, ontologies or metadata, have no semantic meaning: they are experimentally designed

to be of some help for some specific kinds of data, tasks or users. For example, Knowledge Zone (Lewen, Supekar, Noy, & Musen, 2006) allows its users to rate ontologies with numerical or free text values for criteria such as "usage", "coverage", "correctness" and "mappings to other ontologies", also allows its users to rate each other users' ratings, and uses all these ratings to retrieve and rank ontologies. This approach compounds several problems: (i) whole ontologies are rarely genuinely/intuitively comparable (given two randomly selected ontologies, it is very rare that one fully includes or specializes the other), (ii) giving numerical values for such criteria is rather meaningless, (iii) textual values for each of such criteria cannot be automatically organised into a semantic network, (iv) two sets of criteria are rarely comparable (one set rarely includes all the criteria of the other set and has higher values for all these criteria), and (v) similarity measures on criteria only permit to retrieve possibly "related" ontologies: the work of understanding, comparing or merging their statements still has to be (re-)done by each user.

To sum up, however sophisticated, techniques that index resources are inherently limited in their possibilities and usefulness for information seekers. Furthermore, since they do not provide re-use mechanisms, they force information providers to repeat or re-describe information elsewhere described and thus add to the volume of redundant data that information seekers have to sift through. Yet, techniques to index data or people form the bulk of LO retrieval/management techniques and Semantic Web related techniques, for example in the Semantic Learning Web (Stutt & Motta, 2004) and the Educational Semantic Web (Devedzic, 2004). Although the number and apparent variety of these techniques is huge, our definitions permit to categorize most of them as follow:

- As annotation tools permitting their users to index or relate resources or metadata (i) by informal terms (e.g., folksonomy tools and topic map based tools), (ii) by terms from a small predefined small list such as the Dublin Core metadata or argumentation relations as in ScholOnto (Buckingham-Shum, Motta, & Domingue, 1999), (iii) by terms from an informal hierarchy such as the DMOZ topic hierarchy, (iv) by terms from a lexical database such as WordNet, (v) by terms from a semantically organized ontology such as the SUMO, (vi) by terms from an ontology that can be updated by users, as in WebKB-2 (Martin, 2003a), (vii) by attribute-value pairs with textual/numerical values, (viii) by restricted kinds of knowledge representations (e.g., semantic wikis), or (ix) by expressive knowledge representations, as in WebKB-2 which uses Conceptual Graphs and Formalized-English.
- As tools automatically indexing or relating resources or metadata (i) by terms from a predefined small list, (ii) by informal terms automatically organized into a hierarchy via techniques such as Latent Semantic Indexing, Formal Concept analysis or terminological analysis, (iii) by terms from lexical databases via natural language parsing (NLP) techniques, (iv) by attribute-value pairs with textual or numerical values, (v) by a measure of similarity between resources and/or their metadata (vi) by informal sentences (e.g., summarizing tools) using statistical or NLP techniques, or (vii) by restricted kinds of knowledge representations (e.g., question-answering tools which index sentences in documents but are not able to represent most of the semantic content of different sentences and hence organize it) via NLP techniques or ad-hoc Web site wrappers. Shadbolt et al. (2006) acknowledge that current "Semantic Web"-like applications still use ad-hoc wrappers from particular Web documents or databases.

As previously noted, current LO-related standards focus on associating simple metadata to (big parts of) documents, and current LOs are almost never about *one un-decomposable statement only*. For example, a typical LO about Java is an "Introduction to Java" listing some features of Java and giving an example of code, instead of being a relation between Java and one of its features. According to the IEEE LTSC (2001), a LO should consist of 5 to 15 minutes of learning material. Each of such LOs cannot be not a "truly re-usable LO" (object) but is a package of objects selected and ordered to satisfy a certain curriculum. Although such packages are useful for pedagogical purposes and ease the task of most course designers since they are ready-made packages, they are black-box packages, that is, their decomposition into objects from a shared well-organised KB has not be made explicit and hence they cannot be easily modified nor compared or efficiently retrieved: they can only be retrieved via keywords, not via arbitrary complex conceptual queries on the objects they contain or, from a browsing viewpoint or a conceptual querying efficiency viewpoint, they cannot be organized into a lattice (partial order) according to the objects they combine.

2.2. Approaches Based on Either Fully Formal or Mostly Informal Resources Are Not Scalable

Some information repository projects use formal KBs, e.g., the Open GALEN project which created a KB of medical knowledge, the QED Project which aims to build a formal KB of all important, established mathematical knowledge, and the Halo project (Friedland et al., 2004) which has for very long term goal a system capable of teaching much of the world's scientific knowledge by preparing and answering test questions for students according to their knowledge and preferences. 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 70 pages of a chemistry book and had encouraging but far-from-ideal results. Hence, such fully formal KBs are not adequate for scalable information sharing or retrieval.

Informal documents (articles, emails, wikis, etc.), that is, documents mainly written using natural languages such as English, as opposed to knowledge representation languages (KRLs), do not permit objects to be explicitly referred and interconnected by semantic relations. This forces document authors to summarize what has been described elsewhere and make choices about which objects to describe and how: level of detail, presentation order, etc. This makes document writing a time consuming task. Furthermore, the lack of detail often makes difficult for people or softwares to understand the precise semantic relations between objects implicitly referred to within and across documents. This leads to interpretation or understanding problems, and limits the depth and speed of learning since retrieving or comparing precise information has to be done mostly manually. The automatic indexation of sentences within documents permits to retrieve sentences that may contain all or parts of some required information (this process is often called "question answering"; tools supporting it are evaluated by the TREC-9 workbenches) but the lack of formalization in the sentences does not permit to extract and merge their underlying objects and relations.

Cognitive maps and concept maps (Novak, 2004) - or their ISO version, topic maps - have often been used for teaching purposes. However, they are overly permissive and hence do not guide the user into creating a principled, scalable and automatically exploitable semantic network. For example, they can use relations such as "of" and nodes such as "other substances" instead of semantic relations such as "agent" and "subtask", and concept names such as "non_essential_food_nutrient". Thus, concept maps are often more difficult to understand or retrieve, aggregate and exploit than regular informal sentences (from which, unlike deeper representations, they can currently be automatically generated); Sowa (2006) gives commented examples.

Similarly, the modelling of the preferences and knowledge of students or other people is often very poor, e.g., a keyword for each known LO (e.g., "Java") and a learning level for it (e.g., "advanced"). This is for example the case with the CoAKTinG project (Page et al., 2005) which aims to facilitate collaboration and data exchange during or after virtual meetings on a semantic grid, and the Grid-E-Card project (Gouardères, Saber, Nkambou, & Yatchou, 2005) which manages a model of certification for each LO and student on a grid to facilitate her learning and her insertion within relevant communities. A more fine-grained approach in which all the statements for which a student has been successfully tested on are recorded is necessary for efficacy and scalability purposes.

We believe that the main reasons why more knowledge-oriented solutions are not developed can be listed as follow: 1) most people, including many tool developers, have little or no knowledge about semantically explicit structures, 2) many tool developers fear that people will be "scared away" by the looks of such structures or by having to learn some notations, 3) precise and correct knowledge modelling is complex and time-consuming, 4) KB systems are not easy to develop, especially user-friendly ones supporting collaboration between their users, 5) there currently exists a lot of informal legacy data but very little well-organized explicit knowledge.

Point 2 was the reason given used by many creators of "knowledge-oriented" hypermedia systems or repositories for the limited expressiveness of their formal features or notations, e.g., the creators of SYNVIEW (Lowe, 1985), AAA (Schuler & Smith, 1992), ScholOnto (Buckingham-Shum et al., 1999) and the Text Outline project (Sanger, 2006). Shipman and Marshall (1999) note that the restrictions of knowledge-based hypermedia tools often lead people not to use them or to use them in biased ways. Although this fact appears to be presented as an argument against knowledge-based tools, it is actually an argument against the restrictions set to ease the tasks of tool developers (especially for designing graphical interfaces) and supposedly to avoid confusing the users. We agree with the conclusion of Shipman and Marshall (1999) that annotation tools should provide users with generic and expressive structuring features but also convenient default options, and the users should be allowed to describe their knowledge at various levels of details, from totally informal to totally formal so that they can invest time in knowledge representation incrementally, collaboratively and only when they feel that the benefits out-weight the costs.

The above points 1 to 5 are valid but we believe that effective or scalable knowledge sharing and retrieval cannot be achieved without a global virtual KB, and to a large extent, without this KB being collaboratively updated by the information providers. Although this requires the learning of graphical or textual notations for representing information precisely, we will probably not be a problem in the long term: the need for programming languages and workflow/database modelling notations is already well accepted and more and more students learn them. Since the need for small LOs has been recognized and since it is part of the roles of teachers and researchers to (re-)present things in explicit and detailed ways, a global virtual KB is likely to be updated by them first. Their students would then complement it, thus providing their teachers a way to evaluate their knowledge and analytic skills.

2.3. Approaches Based on Mostly Independently Created (Semi-)Formal Resources Are Not Scalable

Like previous distributed knowledge sharing strategies, the W3C's strategy is minimal: the W3C only proposes a low-level KRL (RDF+OWL) and some optional rudimentary "best practices" (Swick et al., 2006), and envisages the Semantic Web to be composed of many small KBs (RDF documents), more or less independently developed and

thus partially redundant, competing and very loosely interconnected since the knowledge provider is expected to select, import, merge and extend other people's KBs into her own (Rousset, 2004). This formal document relying approach has problems that are analogue to those we listed for informal documents: (i) finding relevant KBs, choosing between them and combining them is difficult and sub-optimal even for a knowledge engineer, let alone for softwares, (ii) a knowledge provider cannot simply add one object "at the right place" and is not helped nor guided by a large KB (and a system exploiting it) into providing precise and re-usable objects that complement the already stored objects, and (iii) as opposed to normalized insertions into a shared KB which directly or indirectly guide all other related insertions, creating new ontologies actually increases the amount of poorly interconnected information to search, compare and merge by people or software agents. Most of current Semantic Web related approaches focus on supporting the manual setting or automatic discovery of relations between formal terms from different ontologies. Euzenat, Stuckenschmidt and Yatskevich (2005) gave an evaluation of such tools and concludes that they are quite understandably very imperfect but can be sufficient for certain applications. Euzenat (2005) recognizes the need for the approach we advocate: (semi-)formal KBs letting both 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. Those ideas are further developed in the next section.

3. Main Focus: Approaches for Scalable Knowledge Sharing

This section focuses on techniques to support the only approach that we deem efficient and scalable for knowledge sharing and retrieval on the internet or within large intranets: the collaborative creation of a global virtual well-organized (semi-)formal KB without redundancies nor implicit inconsistencies. This implies techniques supporting (i) knowledge replication between KBs, (ii) collaborative knowledge edition within a KB, (iii) the valuation and filtering of knowledge or knowledge sources, and (iv) knowledge normalization.

3.1. Supporting Knowledge Sharing Between KBs

In a global virtual KB, it should not matter which (non-virtual) KB a user or agent chooses to query or update first. Hence, 1) object additions/updates made in one KB should be replicated into all the other KBs that have a scope 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. To achieve those points, in (Martin, Eboueya, Blumenstein, & Deer, 2006) we note that each KB server can periodically checks more general servers, competing servers and slightly more specialized servers, and (i) integrates all the objects generalizing the objects defined in the "reference collection"[1] that defines the scope of this KB server, (ii) integrates all the objects (and direct relations from/to them) 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), and (iii) also stores the URLs of the direct specializations of the generalizations of the objects in the reference collection (this is needed for any object in the global virtual KB to be directly or indirectly referred to). This seems the simplest approach because (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 or ontology for all the objects is necessary since classifying servers according to fields or domains is far too coarse to index or retrieve knowledge from distributed servers, e.g., knowledge about "neurons" or "hands" are relevant to many domains. This approach would work with servers on the Web but also in a peer-to-peer network where each user has her own KB server: the main difference is that a peer-to-peer network permit to implement systematic push/pull mechanisms instead of relying on KB servers to regularly check KBs of other servers and integrate new additions. We found no other research aiming to solve the above specifications 1 or 2. Works dealing with "Ontology Evolution in Collaborative Environments", e.g., (Vrandecic et al., 2005) and (Noy, Chugh, Liu, & Musen, 2006), or (Rousset, 2004) in a peer-to-peer context, are solely about accepting/rejecting and integra-ting changes made in other KBs, not about making these KBs have an equivalent content for their shared sub-scopes.

Integrating knowledge from other servers of large KBs is not easy but it is easier than integrating dozens or hundreds of (semi-)independently created small KBs. Furthermore, since in our approach the first integration from a server is loss-less, the subsequent integrations from this server are much easier. A more fundamental obstacle to the widespread use of this approach is that many industry-related servers are likely to make it difficult or illegal to mirror their KBs; however, this problem hampers all integration approaches. The above described replication mechanism is a way to combine the advantages commonly attributed to "distributed approaches" and "centralized approaches". The

[1] A reference collection is a list of objects with possibly some maximum depth for some relations from these objects. For a completely general server, this collection is reduced to most general conceptual category imaginable (often named "Thing").

inadequacy of this terminology - and its related misconceptions - are thereby also highlighted: (i) not just "mostly independently created resources" can be distributed, and (ii) as shown by the next two sub-sections, "collaboratively editing a same KB" (i.e., centralization) does not imply that the users have to agree or even discuss terminological issues or beliefs, nor that a committee making content selection or conflict resolution for the users is necessary.

3.2. Supporting Collaborative Knowledge Editions Within a KB

Most knowledge servers support concurrency control and users' permissions on files/KBs but WebKB-2 (Martin, 2003a) is the only server having editing protocols permitting and encouraging people to tightly interconnect their knowledge into a shared KB, without having to discuss and agree on terminology or beliefs, and while keeping the KB consistent. Co4 (Euzenat, 1996) had knowledge sharing protocols based on peer-reviewing for finding consensual knowledge: their output was a hierarchy of KBs, the uppermost ones containing the most consensual knowledge while the lowermost ones were the KBs of the contributing users. All other "protocols" used in knowledge portals (Lausen et al., 2005) or knowledge oriented approaches in peer-to-peer networks (Rousset, 2004) or Semantic Grids (Page et al., 2005) focus on managing the integration of some source KB into a private/shared target KB: these protocols are not guiding nor even permitting the users of the two involved KBs to tightly interconnect their knowledge. The next paragraph summarises the principles of WebKB-2's editing protocols.

Each category identifier is prefixed by an identifier of the category creator (who is also represented by a category and thus may have associated statements). Each (formal or informal) statement also has an associated creator and hence, if it is not a definition, may be considered as a belief. Any object (category or statement) may be re-used by any user within her statements. The removal of an object may only be done by its creator but a user may "correct" a belief by connecting it to another belief via a "corrective relation". Definitions cannot be corrected since they are neither true nor false; a user "fg" is entitled to define fg#cat as a subtype of the WordNet type wn#chair: there is no inconsistency as long as the ways these types are further defined respect the constraints associated to each other. If entering a new belief introduces a redundancy or an inconsistency that is detected by the system, it is rejected. The user may then either correct this belief or re-enter it again but connected by specialization relations (e.g. "example") or "corrective relations" (e.g., "corrective_generalization") to each belief it is redundant or inconsistent with. For example, here is a Formalized-English statement by Joe that corrects an earlier statement by John: `any bird is agent of a flight'(John) has for corrective_specialization `most healthy French birds are able to be agent of a flight' '(Joe). The use of corrective relations allows and makes explicit the disagreement of one user with (her interpretation of) the belief of another user. This also technically removes the cause of the problem: a proposition A may be inconsistent with a proposition B but a belief that "A is a correction of B" is not technically inconsistent with a belief in B. Choices between beliefs may have to be made for an application, but then the explicit relations between beliefs can be exploited, for example by always selecting the most specialized beliefs.

3.3. Supporting the Valuation and Filtering of Knowledge or Knowledge Sources

The above described recording of each object's creator, and the possibility for any user to represent information about each creator, permit to combine conceptual querying "by the content" with conceptual querying "on the creators". For example, WebKB-2 allows any user to set up filters on certain (kinds of) creators to avoid their knowledge being displayed during browsing or within query results. This is handy when bad quality knowledge from certain users becomes a nuisance for exploring and comparing the objects of certain domains despite the conceptual organization of the KB and hence its limited amount of redundancies. However, to allow a much better filtering of knowledge and/or their sources, additional information on each statement and each statement creator need to be recorded and exploited: their originality, popularity, acceptation and other characteristics related to the "usefulness" of a statement or creator. In (Martin et al., 2006), we gave a template algorithm to quantify the usefulness of each statement in a KB, and then also on each of their creators, based on votes from users on statements and on how each statement is (counter-)argued using argumentation relations. To be even more useful, this algorithm should accept parameters permitting each user to specify her own view about which kinds of statements or users should be displayed and, if so, how. This approach eliminates the need for (i) allowing or forcing "special users" to perform some content selection in the KB for other users, thereby restricting the scope, goals and interest of the KB, or (ii) allowing any user to delete anything, as in wikis, which leads to edit wars. However, there is still a need for some special users to remove (or not) completely irrelevant statements (spam) that have been voted as such by some users and not prevented automatically. Given the way our template algorithm attributes a usefulness value to each statement and each user, this approach should incite the users to be careful and precise in their contributions and give arguments for them: unlike in traditional discussions or reviews, a value for each statement can be given by the template algorithm and each user can refine the problematic statements to improve them and be rewarded.

In his description of a "Digital Aristotle", Hillis (2004) describes a "Knowledge Web" to which researchers could add "isolated ideas" and "single explanations" at the right place, and suggests that this Knowledge Web could and should "include the mechanisms for credit assignment, usage tracking, and annotation that the Web lacks" (pp. 4-5), thus supporting a much better re-use and evaluation of the work of a researcher than the current system of article publishing and reviewing. Hillis does not give any indication on such mechanisms but those proposed in this sub-section and the two previous ones seem a good basis. Other valuation and trust propagation mechanisms exist, e.g., those of Lewen et al. (2006) referred to in Section 2.1, but unfortunately (i) they are used on attribute-values representing/indexing the content of whole documents, not on the "usefulness" characteristics of precise statements, and (ii) they generally do not take argumentation relations into account. A primitive and informal version of our statement valuation approach was implemented in SYNVIEW (Lowe, 1985). Finally, we mentioned how Co4 allowed its users to evaluate how consensual their knowledge was.

3.4. Supporting Knowledge Entering and Normalization

To ease the automatic or manual comparison of objects within and between KBs, and hence also their retrieval, these objects should be represented as precisely and uniformly as possible. This implies easing and guiding knowledge entering by providing the users with at least the following supports, all of which should be designed to ease the adoption of knowledge modelling "best practices": 1) for each KB, a large well-organized ontology that integrates the various existing ontologies related to the scope of the KB, 2) knowledge entering/querying/entering interfaces exploiting these ontologies and hence dynamically generated from them, 3) expressive, intuitive and concise KRLs, and 4) parsers for simple natural language sentences that propose normalized representations for these sentences. Many complementary knowledge modelling methodologies (e.g., CommonKADS, Ontoclean, Methondology and On-To-Knowledge) and "best practice" rules exist but most of them are un-supported by all low-level KRLs (e.g., KIF, the Knowledge Interchange Format, and RDF, the Resource Description Format), by almost all other KRLs and ontologies and by most KB editors. Almost all the examples and ontologies related to the Semantic Web, including those provided by the W3C, ignore the lexical, structural and ontological best practices that we collected in (Martin, 2000). Some examples are given in Section 7. Only Point 2 of the above four points is not uncommon in advanced KB systems, as for example in SHAKEN (Chaudhri et al., 2001). CYC provides approximate solutions for the four points: it has a parser of English sentences (Witbrock et al., 2003), it has the biggest existing general KB and CycL (the KRL of CYC) is expressive albeit not very intuitive nor concise. However, CYC does not respect lexical, structural and ontological best practices; for example, because of CyCL, CYC often contains statements based on N-ary relations instead of using more explicit and matchable forms using binary relations. Furthermore, CYC does not store the sources of each object (e.g., its creator or a source in a document and the user that represented it into the KB) and does not have protocols to permit the update of the KB by any Web user.

As a step toward Point 1, we transformed WordNet into a genuine lexical ontology and complemented it with many top-level ontologies (Martin, 2003b) into WebKB-2. We have also begun an ontology of knowledge engineering (Martin & Eboueya, 2007) into WebKB-2 and we shall invite researchers and lecturers in this field to represent their ideas, tools and LOs when such additions will be sufficiently guided by the ontology and WebKB-2 to be made in a scalable manner. This means that we have to represent and organise the main tasks, data structures and technique characteristics in knowledge engineering. An ontology such as the Semantic Web Topics Ontology of ISWC 2006 is by no mean usable for knowledge representation and is not even scalable for document indexation since (i) it does not follow knowledge representation/sharing best practices, is not integrated into a lexical ontology, and updates should be suggested to its creators by email or via a wiki, and (ii) it is based on "topics" and uses quite vague relations such as topic_subtopic, topic_requires, topic_relatedTo and topic_relatedProjects, and hence does not permit the user to find "a right place" to insert a new concept - as noted by Welty and Jenkins (1999), placing a topic into a specialization hierarchy of topics is quite arbitrary, whereas a category for a task or a data structure has a unique correct place into a partOf/specializationOf hierarchy of tasks or data structures, given the intended formal meaning of the categories and the formal meanings of the used partOf/specializationOf relations.

As a step toward Point 3, WebKB-2 proposes notations such as "Formalized English" (FE), "Frame Conceptual Graphs" (FCG) and "For-Links" (FL; a sub-language of FCG when quantifiers need not be used). They are more high-level and compact than currently existing notations and often much more expressive too (Martin, 2002). High-level means intuitive and normalizing: the syntax of our notations includes many components (e.g., various extended quantifiers and collection "interpretations") that (i) would be very difficult for users to define correctly and in comparable or formally exploitable ways, (ii) make the syntax more English-like, and (iii) lead the users to follow best practices and hence provide more precise and automatically comparable knowledge, thus, more retrievable and checkable for redundancies and inconsistencies. More compact means that more knowledge can be displayed in a structured way in a short amount of space, which is very important to ease the manual retrieval and comparison of

knowledge in a large KB. This is one of the reasons why KB systems should allow the entering, querying, display and browsing of knowledge using textual notations in addition to graphic notations. The following tables show examples of simple representations in FE, FCG and FL, languages that we are still extending. RDF translations of them would be long and ad-hoc. We packed many details into these examples and we invite the reader to really delve into these details in order to get a better intuition of the proposed approach.

Table 1. Compact representations of English sentences into FL.

Note. The creators of the terms are not specified and hence the representations are informal.

```
En:  According to the user with identifier "jo", (i) any human body has at most 2 arms and exactly 1 head,
     and (ii) most arms belong to at most 1 human body.
     According to "pm", male_body and female_body are exclusive subtypes of human_body,
     and most human bodies have legs.
     According to "oc", most human_bodies are able to sleep for 12 hours.

FL:  human_body  part:  arm [any->0..2(jo), 0..1<-most(jo)]    head [any->1(jo)]    leg [most->0..*(pm)],
                        subtype:  excl{ male_body(pm) female_body(pm) }(pm),
                        can be agent of:  [(sleep, period: 12 hour)][most->a(oc)];

//this last line could also be wrtiiten in FCG:
     [most human_body can be agent of: (a sleep, period: 12 hour)](oc);
```

Table 2. Formal representations of an English sentence into FL, FCG and KIF.

```
En:   According to "jo", most human_body (as understood in WordNet 1.7) may have for part (as understood by "pm")
       one or two legs (as defined by "fg") and have exactly 1 head (as understood by "oc").

FL:   wn#body   pm#part:   0..2 fg#leg (jo)     1 oc#head (jo);

FE:    `most wn#body  pm#part at most 2 fg#leg and for pm#part 1 oc#head' (jo);

FCG: [most wn#body,  pm#part: at most 2 fg#leg,  pm#part: 1 oc#head](jo);

KIF:   (believer '(forall ((?b wn#body)) (atLeastN 1 '?l  fg#leg     (pm#part '?b ?l )))  jo)
       (believer '(forall ((?b wn#body)) (exactlyN 1 '?h oc#head (pm#part '?b ?h)))  jo)
```

Table 3. Interconnection of semi-formal statements in FL.

Notes. The creators of the statements could have been made explicit. The very last statement is in FE (it begins with an opening single quote). The terms used below for the relations and with an underscore inside are informal but the relevant related formal terms/categories for them can be automatically found). To normalize the formulation of the statements and ease their organization and retrieval, most of the statements begin by a process and all the processes have related formal terms/categories. The parenthesis are used for two different purposes which the indentation help understand: (i) allowing the direct representation of links from the destination of a link, and (ii) representing meta-information on a link, such as its creator (e.g., "pm" or "fg") or a link on this link. Dashes are used for joint arguments/objections (e.g., a rule and its premise). Most notations proposed by argumentation systems do not have this expressiveness and compactness, and hence restrict or bias the work of their users.

```
"knowledge_representation_or_exchange_with_XML is useless"
    argument: ("the use_of_XML_tools_by_KBSs is a useless additional task"
                    argument: "the internal_use_of_XML_by_a_KBS is useless" (pm,
                       objection: "knowledge_representation_or_exchange_with_XML is possible" (fg,
                          objection: "knowledge_representation_or_exchange_with_non-XML-languages is possible" (pm),
                          objection: "knowledge_representation_in_a_KBS_with_a_non-XML_language is necessary" (pm)))
                )(pm);

"knowledge_representation_or_exchange_with_XML is possible"
    argument: - "the re-use_of_a_classic_XML_tool (parser, XSLT, ...) is permitted by the use_of_an_XML_notation" (pm)
              - "the re-use_of_a_classic_XML_tool is possible even when a graph-based model is used" (pm),
    argument of: ("a KR_language should have at least one XML_notation for input/output format",
                       specialization: "the Semantic_Web_KRL should have an XML_notation" (pm),
                       specialization of: `a KR_language can have for notation an XML_notation' (pm)
                  )(pm);
```

We have used FL to represent the content of three courses at Griffith Uni: "Workflow Management", "Systems Analysis & Design", and "Introduction to Multimedia". Figures 1 shows an extract of the input file for the first course, while Figure 1 and Figure 2 show simple queries on its knowledge. Nearly each sentence of each slide for these courses has been represented into a semantic network of tasks, data structures, properties, definitions, etc. The students of these courses have recognised the help that the semantic network provides them in relating and comparing information otherwise scattered in many different slides and other lecture materials. Having to learn FL was however perceived as a problem, especially by the students who were evaluated on their contributions to the semantic network (Martin, 2006). An intuitive table-based knowledge entering/display interface for FL should reduce this problem.

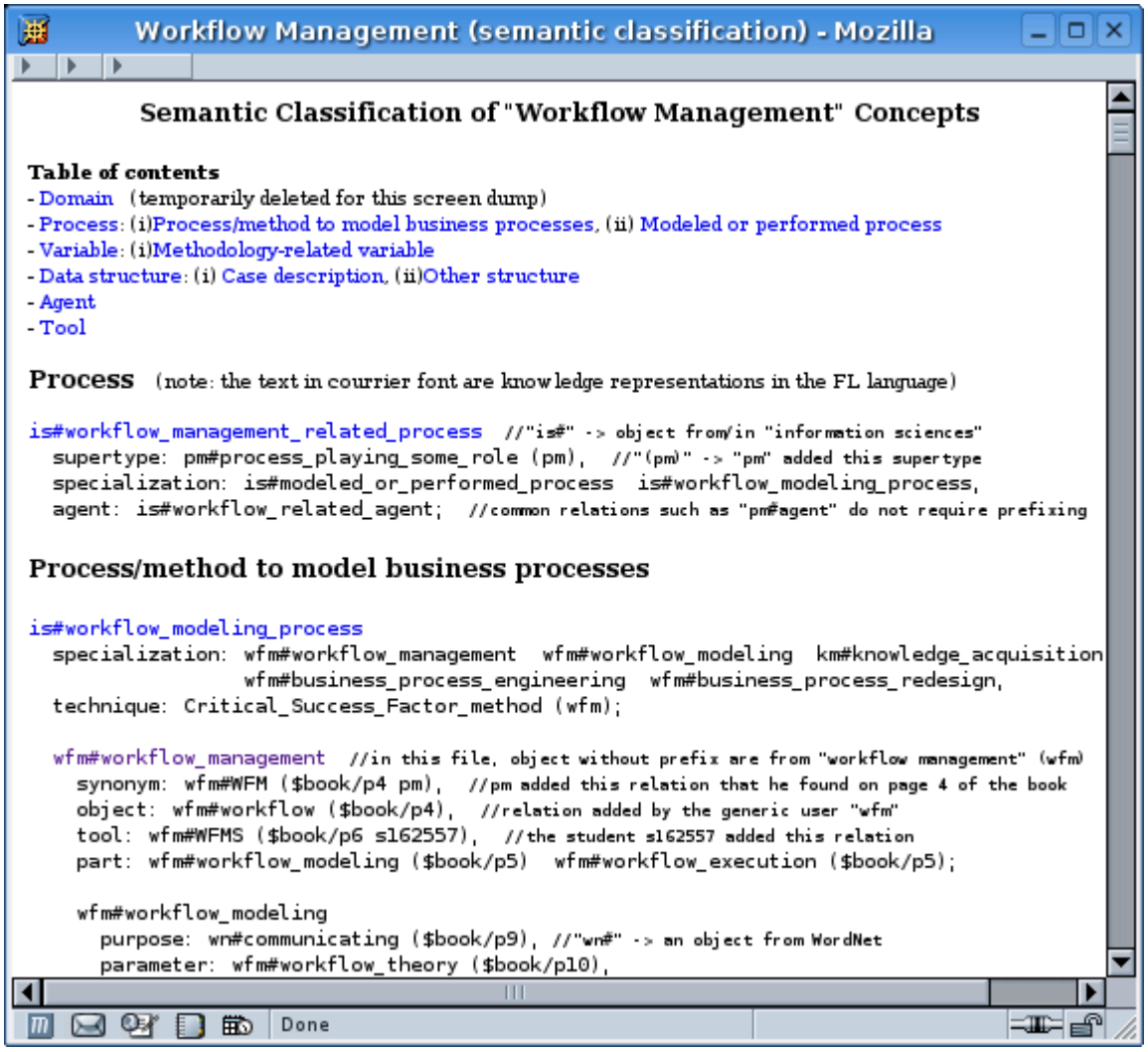

Figure 1. Extract from a file representing statements from a book in Workflow Management (book referenced here by the variable $book; any Web user can create such a file and ask WebKB-2 to parse it and hence integrate its knowledge representations into the shared KB).

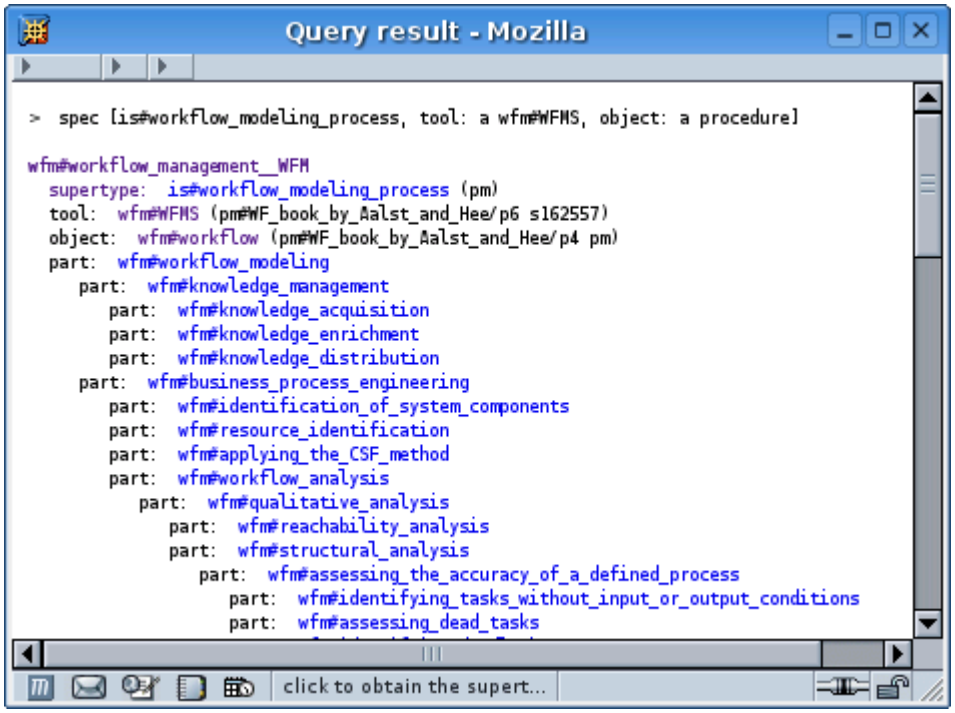

Figure 2. A search for the *spec*ializations of a statement in FCG and its first result (clicking on wfm#workflow_management returns the same result, here displayed in an informal format looking like FL).

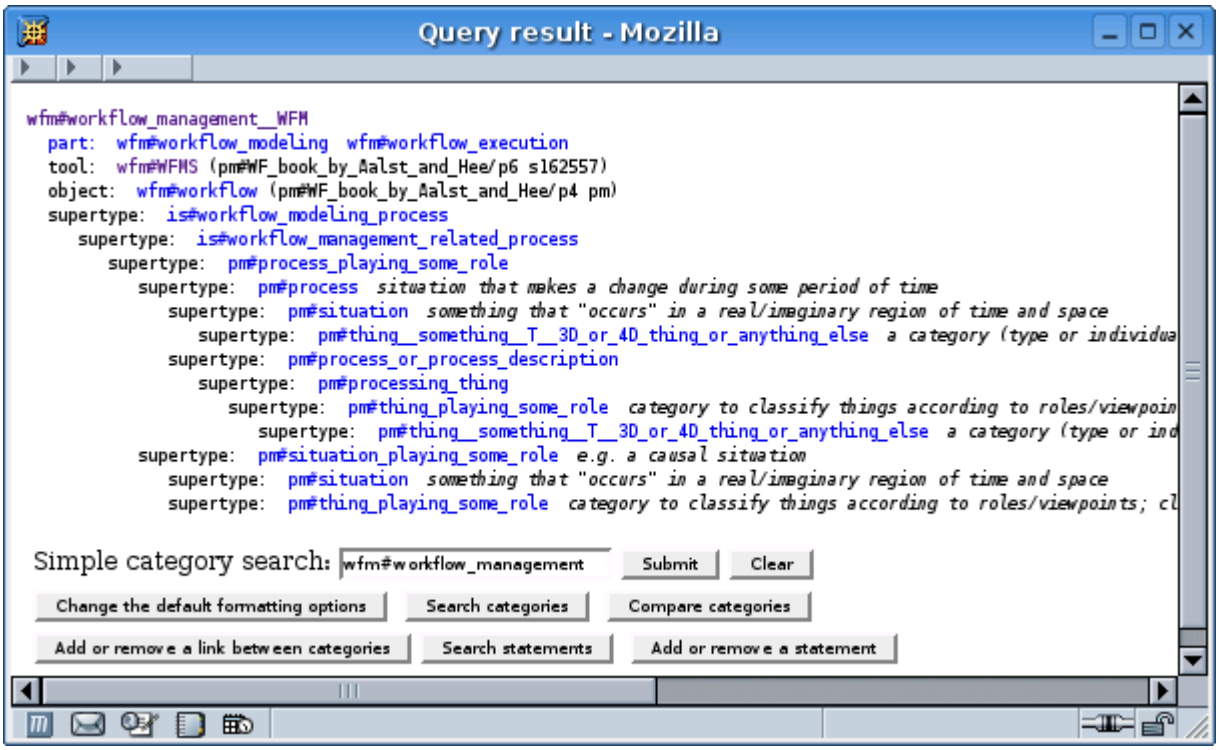

Figure 3. Expansion of the supertypes of wfm#workflow_management.

4. Future Trends: Bigger and Fewer Knowledge Repositories

Nowadays, many businesses grow or merge to stay competitive, and de-facto standards tend to persist despite their widely recognized shortcomings, especially in information technology. The KB and knowledge sharing conventions or mechanisms of the first company that will propose a general KB that people will be able to update in a somewhat organized way are likely to quickly become de-facto standards in the same way that the Web, Google and Wikipedia quickly became widely used. Given current knowledge sharing practices, it is unfortunately unlikely that this initial KB and chosen conventions or mechanisms will be the best ones for scalability purposes. In any case, this KB will be collaboratively updated by all kinds of persons (researchers, lecturers, students, company employees, etc.) and purposes (storing LOs, advertising or giving feedbacks on products, etc.). Indeed, we have shown that a KB server can be used by many people for collaboratively organizing and valuating knowledge at various levels of details, and that alternative technologies are less efficient for sharing and retrieving information.

One hypothesis behind our approach is that a sufficient number of persons will take the time to be precise and learn notations and conventions to do that. We do not think this will be a problem once the approach becomes popular with researchers, teachers and students, and we concluded in Section 2.2 that this was likely to happen. The social success of Wikipedia shows that despite its problems many persons are willing to contribute, and our approach would solve these problems. In this approach people can engage in "structured discussions" by connecting statements via argumentation/corrective relations, thereby not only representing debates in unprecedentedly structured ways but are also collaboratively evaluating themselves on each of their statements; this intellectual challenge and opportunity for recognition may attract a lot of people. 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 approaches and Brandom's model of discursive practice (Brandom, 1998).

5. Conclusion

We argued that a virtual global normalised well-organized collaboratively-updated formal and semi-formal KB is necessary and achievable for the scalable and efficient sharing and retrieval or comparison of precision-oriented kinds of information (LOs included) within intranets or on the internet, and therefore as a shared medium for the tasks of publishing, researching, teaching, learning, annotating, evaluating and collaborating. In comparison, synchronous approaches (e.g., on-line chats and face-to-face teaching) and approaches based on indexing or relating formal or informal documents or KBs, are extremely sub-optimal for information publishing, retrieval, comparison and learning. Ideally, a normalized KB is like a decision tree: the place or way to insert or find information is quickly found, however huge the KB, and the existing information (fact, hypothesis, feedback, etc.) can be incrementally completed or refined. Documents often do not contain precise enough information to create such a KB directly from them; the proposed approach leads information providers to deepen and structure their knowledge and permits to evaluate or filter out each of the individual contributions. Automatic knowledge extraction, alignment or merging methods are needed to help building this KB but need to be adapted to take into account knowledge sharing best practices and used for combining the advantages of centralisation and distribution rather than just creating new resources. Documents and synchronous collaboration or teaching will always exist and be needed but these works will hopefully also lead to the completion of more semantically structured media and hence permit other people to easily find and re-use the results of these works.

References

Brandom, R. (1998). Action, norms, and practical reasoning. *Noûs*, 32(12), 127-139, Pittsburgh, PA: Blackwell Publishing.

Buckingham-Shum, S., Motta, E., & Domingue, J. (1999). Representing scholarly claims in internet digital libraries: A knowledge modelling approach. In S. Abiteboul & A.-M. Vercoustre (Eds.), *European Conference on Digital Libraries* (pp. 423-442), Paris, France: Springer-Verlag.

Chaudhri, V., Rodriguez, A., Thoméré, J., Mishra, S., Gil, Y., Hayes, P., et al. (2001). Knowledge entry as the graphical assembly of components. *K-Cap'01* (pp. 22-29), BC, Canada: ACM Press.

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

Downes, S. (2001). Learning objects: resources for distance education Worldwide. *International Review of Research in Open and Distance Learning*, 2(1).

Euzenat, J. (1996). Corporate memory through cooperative creation of knowledge bases and hyper-documents.. In B. Gaines (Ed.), *Knowledge Acquisition Workshop* (pp. 1-18), Banff, CA, Canada: ksi.cpsc.ucalgary.ca/KAW

Euzenat, J., Stuckenschmidt, H., & Yatskevich, M. (2005). Introduction to the Ontology Alignment Evaluation 2005. In B. Ashpole, M. Ehrig, J. Euzenat & H. Stuckenschmidt (Eds.), *K-Cap'05* (pp. 61-71) Banff, Canada: CEUR-WS.org.

Euzenat, J. (2005). Alignment infrastructure for ontology mediation and other applications. In M. Hepp, A. Polleres, F. Harmelen & M. Genesereth (Eds.), *International workshop on Mediation in semantic web services* (pp. 81-95), Amsterdam, Netherlands: CEUR-WS.org.

IEEE LTSC (2001). IEEE learning technology standards committee glossary. *IEEE P1484.3 GLOSSARY WORKING GROUP, draft standard 2001.*

Friedland, N.S, Allen, P., Mathews, G., Witbrock, M., Baxter, D., Curtis, et al. (2004). Project Halo: Towards a digital aristotle. *AI Magazine*, 25(4), 29-48.

Gouardères, G., Saber, M., Nkambou, R., & Yatchou, R. (2005). The Grid-E-Card: Architecture to share collective intelligence on the grid. *Applied Artificial Intelligence*, 19(9-10), 1043-1073.

Hillis, W.D. (2004). "Aristotle" (the knowledge web). *Edge Foundation, Inc.*, 38.

Hodgins, W. (2006). Out of the past and into the future: Standards for technology enhanced learning. In U. Ehlers and J. Pawlowski (Eds.), *Handbook on Quality and Standardisation in E-Learning* (pp. 309-327). Springer.

Horn, R. (2003). *Mapping great debates: can computers think?*. Retrieved July 2, 2007, from http://www.macrovu.com/CCTGeneralInfo.html

Lausen, H., Ding, Y., Stollberg, M., Fensel, D., Lara, R., & Han, S. (2005). Semantic web portals: state-of-the-art survey. *Journal of Knowledge Management*, 9(5), 40-49.

Lewen, H., Supekar, K.S., Noy, N.F., & Musen M.A. (2006). Topic-specific trust and open rating systems: An approach for ontology evaluation. *Workshop on Evaluation of Ontologies for the Web* at WWW'06, Edinburgh, UK: km.aifb.uni-karlsruhe.de/ws/eon2006

Lowe, D. (1985). Co-operative Structuring of Information: The representation of reasoning and debate. *International Journal of Man-Machine Studies*, 23(2), 97-111.

Martin, P. (2000). Conventions and notations for knowledge representation and retrieval. In B. Ganter & G.W. Mineau (Eds.), *International Conference on Conceptual Structures* (pp. 41-54), LNAI 1867, Springer.

Martin, P. (2002). Knowledge representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English. In U. Priss, D. Corbett & G. Angelova (Eds.), *International Conference on Conceptual Structures* (pp. 77-91), LNAI 2393, Springer.

Martin, P. (2003a). Knowledge representation, sharing and retrieval on the web. In N. Zhong, J. Liu & Y. Yao (Eds.), *Web Intelligence* (pp. 263-297), Springer.

Martin, P. (2003b). Correction and extension of WordNet 1.7. In G. Ellis & G. Mann (Eds.), *International Conference on Conceptual Structures* (pp. 160-173), LNAI 2746, Springer.

Martin, P., Eboueya, M., Blumenstein, M., & Deer, P. (2006). A network of semantically structured wikipedia to bind information. In T. Reeves & S. Yamashita (Eds.), *World Conference on E-Learning* (pp. 1684-1702), Honolulu, HI: AACE.

Martin, P. (2006). Griffith e-learning fellowship report. Retrieved July 2, 2007, from http://www.webkb.org/doc/papers/GEL06/

Martin, P., & Eboueya, M. (2007). Sharing and comparing information about knowledge engineering. *WSEAS Transactions on Information Science and Applications*, 5(4), 1089-1096.

Noy, N.F., Chugh, A., Liu, W., & Musen, M. A. (2006). A framework for ontology evolution in collaborative environments. In I. Cruz, et al. (Eds.), *International Semantic Web Conference* (pp. 544-558), LNCS 4273, Springer.

Novak. J.D. (2004). Reflections on a half century of thinking in science education and research: Implications from a twelve-year longitudinal study of children's learning. *Canadian Journal of Science, Mathematics, and Technology Education*, 4(1), 23-41.

Page, K., Michaelides, D., Buckingham-Shum, S., Chen-Burger, Y., Dalton, J., De Roure, et al. (2005). Collaboration in the semantic grid: a basis for e-learning. *Journal of Applied Artificial Intelligence*, 19(9-10), 881-904.

Rousset, M-C. (2004). Small can be beautiful in the semantic web. In S. McIlraith, D. Plexousakis, & F. Harmelen (Eds.), *International Semantic Web Conference* (pp. 6-16), LNCS 4273, Springer.

Sanger, L.M. (2006) The future of free information. *Digital Universe Electronic Journal*, 2006-01. Retrieved July 2, 2007, from http://www.dufoundation.org/downloads/Article_2006_01.pdf

Shadbolt, N., Berners-Lee, T., & Hall, W. (2006). The semantic web revisited. *IEEE Intelligent Systems*, 21(3), 96-101.

Schuler, W., & Smith, J.B. (1992). Author's Argumentation Assistant (AAA): A hypertext-based authoring tool for argumentative texts. *Hypertext: concepts, systems and applications,* 137-151, Cambridge University Press.

Shipman, F.M., & Marshall, C.C. (1999). Formality considered harmful: experiences, emerging themes, and directions on the use of formal representations in interactive systems. *Computer Supported Cooperative Work*, 8, 333-352.

Swick, R., Schreiber, G., & Wood, D. (2006) Semantic web best practices and deployment working group. Retrieved July 2, 2007, from http://www.w3.org/2001/sw/BestPractices/

Sowa, J.F. (2006). Concept mapping. Retrieved July 2, 2007, from http://www.jfsowa.com/talks/cmapping.pdf

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

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

Welty, C.A. & Jenkins, J. (1999). Formal ontology for subject. *Journal of Knowledge and Data Engineering*, 31(2), 155-182.

Witbrock, M., Baxter, D., Curtis, J., Schneider, D., Kahlert, R., Miraglia, P., et al. (2003). An interactive dialogue system for knowledge acquisition in Cyc. In G. Gottlob & T. Walsh (Eds.), *International Joint Conference on Artificial Intelligence* (pp. 138-145), Acapulco, Mexico: Morgan Kaufmann.

7. Key Terms and Their Definitions

Although classic string-matching methods can also be used for retrieving knowledge, **knowledge retrieval** mainly refers to a "conceptual search" or "search by the content", that is, to manual navigation along conceptual relations between objects, or to queries that exploit the formal definitions of these relations. Both cases rely on ***comparisons*** between ***objects (categories or formal/informal statements)***. Two objects are ***incomparable*** when no generalization relation between them has been set manually or can be inferred.

Knowledge normalization aims to ease manual or automatic knowledge comparison and retrieval by reducing the number of incomparable ways information is or can be written and by improving the way objects are (re-)presented and connected. ***Lexical normalization*** involves following object naming rules such as "use English singular nouns or nominal expressions" and "follow the undescore-based style instead of the Intercap style". ***Structural and ontological normalization*** involves following rules such as "when introducing an object into an ontology, relate it to all its already represented direct generalizations, specializations, components and containers", "use subtypeOf relations instead of or in addition to instanceOf relations when both cases are possible", "avoid the use of non binary relations" and "do not represent processes via relations". For example, the above rules lead to the introduction of the concept type "sitting_down" instead of the relation types "sits", "sitsOn" and "sits_on_atPointInTime" which are incomparable. Then, the sentence "some animal sits above some artifact" can be represented in the following explicit form in the Formalized-English notation: "some animal is agent of a sitting_down above some artifact" (this sentence uses the very common basic relations "agent" and "above"). As this example illustrates, knowledge normalization means reducing redundancies as well as increasing the precision and scalability of knowledge modelling. ***Scalable*** knowledge modelling and sharing approaches maintain the possibility of efficiently and correctly finding and/or inserting a piece of information even when the KB become very large. Scalability implies the exploitation of automatic procedures for (i) discovering consistencies and redundancies during knowledge updates, and (ii) filtering knowledge according to various criteria during searches.

Knowledge sharing is the act of publishing information in a more or less normalized way.