Nimis P. L., Vignes Lebbe R. (eds.)
Tools for Identifying Biodiversity: Progress and Problems – pp. 65-70.
ISBN 978-88-8303-295-0. EUT, 2010.

# Learning, Identifying, Sharing

Philippe A. Martin, Noël Conruyt, David Grosser

**Abstract** — This article argues that a cooperatively-built, well-organized, shared knowledge base is a new – and, from certain viewpoints, optimal – kind of support (refining and integrating other kinds of supports) for three complementary tasks: learning about living entities (and how to identify them), supporting their identification, and sharing knowledge about them. This article gives the ideas behind our prototype, and argues that knowledge providers can be not solely specialists, but also amateurs. In essence, for these three tasks, it argues for the (re-)use of much more semantically organized and interconnected versions of semantic wikis or scratchpads.

**Index Terms** — identifying, knowledge sharing, learning, ontologies, semantic wikis.

## 1 Introduction

Current supports for learning about – and identifying – living entities, e.g., the supports listed by the *KeyToNature* project (www.keytonature.eu), are mostly static files (texts, images, …) and tools based on a formal[1] knowledge base (KB*)*. Few tools allow their users to contribute annotations or other information to their formal or informal KB, let alone use them for i) helping identification or learning, and ii) publishing them in a way usable by other tools. Scratchpads [1] and, more generally, semantic wikis[2], allow the cooperative edition and semantic linking of information by any web user, but not in an organized or formal enough way to be re-used by an identification tool or a problem-solving tool, nor to permit the automatic detection of partially redundant/inconsistent information within or between wikis. This automatic detection is essential to permit the semi-automatic and cooperative organization of knowledge into a ***unique semantic network*** and thus permit i) ***scalable information retrieval, comparison, sharing and exploitation***, and hence ii) an ***easier understanding or learning (by amateurs or specialists)*** of the stored information and viewpoints of their authors. Section 2 quickly compares the various current kinds of supports for the learning and sharing of information

*The authors are with the IREMIA laboratory, University of La Réunion. E-mail: (Philippe.Martin, Noel.Conruyt, David.Grosser)@univ-reunion.fr.*

[1] In this article, “formal” means *machine processable and logic-based*, while “semantic” means *formal and organized by semantic relations, e.g., “subtype of”, “physical_part of”, “agent of” and “duration of”.*

[2] Semantic wikis are collaboratively-built documents with some parts indexed by semantic categories or interconnected by semantic relations. See http://semwiki.org *for more details.*

about living entities and hence for helping their identification.

Section 3 introduces elements required to support an approach leading to a ***global KB*** composed of collaboratively-built KBs that have no *implicit*[3] “automatically detectable partial redundancies or inconsistencies” neither within nor between the KBs. As suggested in Section 2, such a global KB – and hence this approach (which is complementary to the other approaches) – is the most useful one from a knowledge-sharing, retrieval and learning viewpoint, but its disadvantages are that i) it requires the users to learn how to read a textual or graphic notation for representing or interconnecting knowledge, and ii) for each domain that has not yet been well represented in the shared KB, the first knowledge providers have a lot of work to do for organizing the information resulting from the use of other approaches. However, this can be done incrementally, whenever the benefits finally becomes clearer than the costs. The elements of this approach are fully or partially implemented in our knowledge server WebKB-2 [2] (webkb.org).

## 2 Quick Comparison of Approaches

The ***smaller*** the sources of information used for knowledge sharing – i.e. the less objects of information (e.g., statements or images) these resources contain – and the ***less contextual*** (hence more explicit, precise and formal) these objects are, the easier it is to automatically index these resources precisely, to filter out the redundancies and to relate these resources via semantic relations, e.g., to organize them into a specialization hierarchy[4]. Then, the ***easier*** it is to ***retrieve*** these resources (by querying or browsing)[5], ***compare*** them (hence, understand and memorize them), ***combine*** them and, more generally, ***exploit*** them for various purposes, e.g., guiding ***identification***. As illustrated in the following paragraphs, these rather obvious ideas are generally well accepted, but their ***ultimate conclusion*** is socially and technically difficult to bring about and hence not directly studied. The conclusion is: there should *ideally* be one and only ***one global semantic network*** (i.e., each index or symbolic resource should contain only one statement or one formal term; in other words, there should be no difference between should symbolic data and meta-data) and, in this network, all manually or automatically detected partial redundancies or inconsistencies are made explicit via semantic relations. In this article, such a global semantic network is called a ***global cbwoKB*** (cooperatively-built well-organized KB).

The ***Learning object*** (LO) related community and standards (e.g., IEEE LTSC)

[3] In this article, implicit means “not made explicit via a semantic relation”.

[4] Related small *individual* statements can often be organized into a specialization hierarchy or an inclusion hierarchy but *sets* of related statements rarely can (the bigger the sets, the less likely).

[5] For example, if the query is of the kind “what are the resources/tools/methods to do ...”, the answer can be a part/subtask/specialization hierarchy (with associated argumentation structures). Such semantically structured answers allow a user to find and compare all relevant objects instead of getting a long list of partially redundant objects or files where original/precise ones are hidden among/behind objects that are more general, more mainstream or from big organizations.

advocate the use of small non-contextual LOs but still only considers the use of static informal documents indexed by keywords. Semantic LO repositories [3] use formal terms or statements for indices. This is also the approach used by STERNA [4].

As highlighted in [5] and [6], the ***Semantic Web*** (SW) community currently essentially focuses on inference mechanisms, KB editors, semantic wikis, social networks, workflow-based cooperation, and the semi-automatic partial interconnection of the content of (semi-)independently created KBs or formal files. Tools created by this community do not *directly* support the creation of a cbwoKB (global or local) and, in a sense, they participate to the problems they are trying to solve since their outputs create new files that are partially redundant or inconsistent with their input files and without semantic relations to make this explicit. The current focus of the SW community is to work with approaches hiding the knowledge representations from the users as much as possible. The problem is then that the semantic network cannot be completed in a meaningful way by the users (only low quality knowledge can be automatically extracted and exploited) nor even browsed to find information. As an example, ***semantic wikis*** are still mainly poorly organized informal documents. Instead, in WebKB-2 the semantic network can be edited by all Web users via cooperation protocols and can be viewed in a more or less structured way via various relatively intuitive syntaxes [7]: Formalized-English, For-Links, etc. However, reading these syntaxes requires a short training and writing knowledge requires the following of some given conventions or “best practices”.

***Scratchpads*** are kinds of semantic wikis which, according to some of their documentation [8], are “independent and unconnected, allowing communities to create distinct customized sites tailored to their needs”. This strongly reduces the possibilities of (semi-)automatically comparing and integrating the content of different scratchpads, and hence works against the goals of identification-related projects like ViBRANT [9] which is based on the use of scratchpads. With a cbwoKB, tailoring can be done by *each* user using filters and presentation rules.

Many identification-related projects use ***databases***, e.g., FishBase (fishbase.org) and Pl@ntNet (plantnet-project.org). They have a regular structure but a rather flat one and users cannot directly contribute to the database: annotations, new objects, new tables (classes of objects), new attributes (relations from/to objects), etc. Finally, the semantics of the objects of these databases is unknown unless their semantic relations to other objects from the Semantic Web are described in a formal file.

Except for WebKB-2, current ***KB servers/editors*** (e.g., Ontolingua, OntoWeb, Ontosaurus, Freebase, CYC and semantic wiki servers) have no shared KB editing protocols and hence either i) let every authorized user modify what other ones have entered (this discourages information entering or leads to edit wars), or ii) require all/some users to approve or not changes made in the KB, possibly via a workflow system (this is bothersome for the evaluators, may force them to make arbitrary selections, and this is a bottleneck to information sharing that often discourages information providers). To complement the generic “knowledge sharing” features of ***WebKB-2*** with identification features, its integration with ***IKBS*** [10], a KB based identification tool, has begun.

## 3 Underlying Ideas of Solutions for the Proposed Approach

To be a ***generic "knowledge sharing"*** support, the shared KB of WebKB-2 has been initialized via a ***loss-less merge of many ontologies*** (sets of formal terms with their associated definitions/constraints/inter-relations): top-level ones (including methodological ones such as DOLCE) and a lexical one (an extension and correction of WordNet) [11]. Knowledge ***normalization rules*** have been collected and extended; simultaneously, various complementary, ***expressive and relatively intuitive notations*** enforcing these rules have been designed [7]. Finally, knowledge sharing protocols have been designed [2]. The ***protocols for the collaborative edition of a shared cbwoKB*** have been implemented and are introduced in the second next paragraph. This is not yet the case for the ***protocols permitting to create a global cbwoKB*** composed of several cbwoKB servers. Their underlying idea is that each of these servers must i) publish its commitment to be a "nexus" for one or several formal terms, that is, to store all information directly related to these terms, and ii) point to other nexus for terms it is not the nexus of. In this way, via redirections of queries and replications of knowledge between servers, it does not matter which server a user updates or queries first, and the advantages of distribution and centralization are thus combined.

WebKB-2 has an ***expressive language model*** (1st-order logic, n-order types, meta-statements and collections) but has a ***simple data model*** since it is built on top of an object-oriented DBMS with only three tables: Term, Relation and Source. Every object of the KB is either a formal/informal term or a formal/semi-formal/informal statement (e.g., a relation between two quantified terms, and a relation on a relation in order to represent some spatial and temporal context). Every object has one or several associated sources: i) the user who created the object, ii) the original resource (e.g., a person, a language, a document) from which the user read/heard/took the object and hence *interpreted* it, and iii) other users who also believe in that object (if it is a statement). ***Lexical conflicts*** are avoided by prefixing formal terms with the identifier of their creators, e.g., wn#bird refers to the most common concept (i.e., meaning) proposed by WordNet for the word "bird".

The next sentences introduce the most important basic ideas behind the shared KB editing protocols of WebKB-2 and hence behind the ways ***semantic conflicts are avoided and the KB kept "well organized"***. A user can re-use any object (term or statement) but can only modify or remove an object that he has created. *Adding, modifying or removing a term* is done by adding, modifying or removing at least one statement (generally, one relation) that uses this term*. A new term can only be added by specializing another term.* Each object must be connected to at least another object via relations of specialization/generalization, identity and/or argumentation (and as many as possible of such relations should be used). If a user adds, modifies or removes a statement (definition or belief) and this creates a *detected conflict* (*redundancy and inconsistency) with another of his statements*, the action is *rejected*. If *adding, modifying or removing a (definition of) a term introduces a conflict with statements of other users,* this conflict highlights an over-interpretation of the term by these other users and

this is automatically *solved by "cloning" the term,* i.e., creating a slightly more general copy of this term for these other users to repair the over-interpretation. If *adding, modifying or removing a belief introduces a detected potential conflict* (partial/total inconsistency or redundancy) involving beliefs created by other creators, it is *rejected*. However, a user may still represent his belief (say, b1) – and thus "loss-less correct" another user's belief that he does not believe in (say, b2) – by connecting b1 to b2 via a corrective relation. E.g., here is a Formalized-English statement by u2 which corrects a statement made earlier by u1:
*u2#` u1#`every bird is agent of a flight´ has for corrective_restriction u2#`most healthy flying_bird are able to be agent of a flight´*.

This statement means: "according to u2, u1's belief that 'every bird flies' is false and a more precise statement is 'most healthy flying birds (the carinates) are able to fly". This way the KB is kept organized and ***then, if necessary, an inference engine can choose between such statements according to the constraints of a particular application***, e.g., it can always choose the most precise version or it can choose the one authored by someone represented as an expert in a certain domain. ***Similarly***, in the same way he creates queries, ***a user can create filters on the content, authors, ..., and popularity of statements*** in order to see *only what he wants to see* when browsing the KB. With this approach, ***every author can represent his beliefs, no selection committee is required, and knowledge integration is loss-less*** (the sources can be regenerated). This approach also ***avoids the problems related to version control or truth-maintenance***.

## 4 Conclusion

This article compared various knowledge sharing approaches and introduced elements necessary to support the most precision-oriented and end-user-controlled approach and the one that combines the advantages of the centralization and distribution. Thus, it is the approach that most permits to i) retrieve and compare knowledge about a living entity and hence learn about it, ii) integrate knowledge from everyone (specialists and amateurs), and iii) leads to create knowledge that directly or indirectly can be re-used by tools to guide identification. Most of these elements are implemented in WebKB-2. It will soon be used to enable Web users to extend the content of FishBase and Pl@ntNet.

## References

[1] V. S. Smith, S. D. Rycroft, K. T. Harman, B. Scott and D. Roberts, ''Scratchpads: a data-publishing framework to build, share and manage information on the diversity of life,'' *BMC Bioinformatics 2009,* 10 (suppl. 14). See also http://scratchpads.eu, 2010.

[2] P. Martin, "Protocols for Governance-free Loss-less Well-organized Knowledge Sharing", *ECAI 2010 workshop on Intelligent Engineering Techniques for Knowledge Bases (I-KBET 2010)*, Lisbon, Portugal, 17 August 2010.

[3] J. S. Carrion, E. G. Gordo and S Sanchez-Alonso, "Semantic learning object repositories", *International Journal of Continuing Engineering Education and Life Long Learning*, vol. 17, 6, pp. 432-446, 2007.

[4] STERNA, "Semantic Web-based Thematic European Reference Network Application", http://www.sterna-net.eu, 2010.

[5] N. Shadbolt, T. Berners-Lee and W. Hall, “The semantic web revisited”, *IEEE Intelligent Systems*, 21, vol. 3, pp. 96-101, May/June 2006.

[6] R. Palma, P. Haase, Y. Wang and R. d’Aquin, “Propagation models and strategies”, *Deliverable 1.3.1 of NeOn* - Lifecycle Support for Networked Ontologies; NEON EU-IST-2005-027595, 2006.

[7] P. Martin, “Knowledge representation in CGLF, CGIF, KIF, Frame-CG and Formalized-English”, *Proc. of ICCS 2002*, Springer LNAI 2393, pp. 77-91, 2002.

[8] P. Martin, “Protocols for Governance-free Loss-less Well-organized Knowledge Sharing”, *Proc. ECAI 2010 workshop on Intelligent Engineering Techniques for Knowledge Bases* (I-KBET 2010), Lisbon, Portugal, 17 August 2010.

[9] ViBRANT, “Virtual Biodiversity Research and Access Network for Taxonomy”, E.U. FP6 project, http://vbrant.org, 2010.

[10] N. Conruyt and D. Grosser, “Knowledge management in environmental sciences with IKBS: application to Systematics of Corals of the Mascarene Archipelago”, *Selected Contributions in Data Analysis and Classification*, Springer Series: Studies in Classification, Data Analysis and Knowledge Organization, pp. 333-344, 2007.

[11] P. Martin, “Correction and extension of WordNet 1.7”, *Proc. of ICCS*, Springer LNAI 2746, pp. 160-173, 2003.