SemWeb

An environment for integrated access to distributed ontological and lexical knowledge bases and their collaborative development and maintenance. A proposal

Dagobert Soergel
College of Library and Information Services
University of Maryland, College Park
Tel. 301-405-2037 Fax 301-314-9145 Home 703-823-2840
ds52@umail.umd.edu

Abstract

This paper presents a proposal and a call for participation in the long-range development of an open, multifunctional, multilingual system for integrated access to many kinds of ontological and lexical knowledge that would support, among many other functions, software localization and CASE tool databases for multinational development teams. A SemWeb system would provide integrated access to existing knowledge bases through a common interface that would search several knowledge bases and collate the data into a common format defined by the SemWeb template. It would also allow the incremental development of a common integrated distributed knowledge base to be shared by a work group or world-wide. A common knowledge base would support collaboration in ontological and lexical projects. Over time, it could absorb data from other knowledge bases, allowing for tighter integration than common interface access. The system would be usable by many levels of users.

1 Introduction

Efficient design and implementation of the user interface, the help system, and of documentation, especially their localization - their expression in the local language - requires access to knowledge about the ontology of the domain and corresponding terminology in multiple languages (Karkaletsis 1995). The same information is needed for many other purposes — writing third-party software books in multiple languages, understanding non-localized software, retrieval of software for the end user or for software reuse, data element standardization, and CASE tools. Some of the information needed is available in existing machine-readable or paper-based repositories, but specialized information must be added to these tools. A multifunctional extensible system that builds on existing resources and is accessible to many users has the highest pay off.

This paper presents a proposal and a call for participation in the long-range development of an open, multifunctional, extensible, multilingual system for integrated access to many kinds of knowledge about concepts and terminology. The system would draw on existing knowledge bases, allow the user to add her own local information, and provide a platform for collaborative incremental development of a common integrated distributed knowledge base. The scope of such a collaboration could be a company, a group of software developers in a given domain, or worldwide and public. Thus there could be common knowledge bases at several levels, and some of the information in a knowledge base could be kept proprietary, accessible only to authorized users. SemWeb needs a software component that manages integrated access to multiple ontological and lexical knowledge bases and an organizational component — the formation of groups of collaborators, from small groups to a world-wide system.

The multiplicity of knowledge bases — existing or created in a SemWeb framework — would be accessed through a common interface that would search several knowledge bases and collate the data into a common format, ready for whatever application. Furthermore, a common integrated distributed knowledge base would provide an environment in which many contributors could carry out ontological and lexical projects more efficiently. Over time, data from other knowledge bases could be incorporated into a common knowledge base through linking or copying, always giving due credit to the source. Either way, such incorporation requires intellectual work but allows for tighter integration than common interface access to multiple knowledge bases. The system would be designed to be usable by many levels of users.

2 Rationale

There are now a multiplicity of order systems/ classifications/ontologies, lexical knowledge bases, linguistic dictionaries, and data element dictionaries, each serving a limited purpose but widely overlapping in their content. The proposed common interface would provide simultaneous access to multiple knowledge bases through a network (the Internet or an intranet) with limited integration of information "on the fly"; going a step further, an integrated distributed SemWeb knowledge base would provide a home for a number of knowledge bases, eliminating duplication while preserving the integrity of each source and establishing relationships across sources. The system would serve many functions and thus justify the major investment it requires. It might appear that this proposal is overly ambitious and that serving so many functions at once is impractical. However, less and less information needs to be added for each additional function, and it is precisely the multifunctionality that makes the considerable investment pay off. The creation of a public SemWeb system should be supported by the public sector as an investment into infrastructure. A common SemWeb knowledge base at whatever level would lead to a savings in development effort and a potentiation of usefulness through the assembly of rich information from many sources that complement each other and through establishing relationships among the concepts and terms from different sources.

Savings in development effort. Much effort is being expended in developing individual ontological and lexical knowledge bases limited by subject, application, and language. These individual knowledge bases overlap considerably; thus there is much duplication of development effort which would be saved in a SemWeb environment.

Potentiation of usefulness. An integrated knowledge base provides rich information on a concept or term, well beyond any specialized system. It provides linkages across scientific and scholarly disciplines. It explicates fine differences of meaning across languages and cultures. It makes conceptual structures that are explicit in one language available to users in other languages. Its usefulness transcends the sum of its parts.

3 Functions

The emphasis or this paper is on the software context with the functions summarized in Figure 1. But to achieve the full benefits, the system should serve the much broader range of functions outlined in Figure 2. The proposed SemWeb system could be applied either way.

4 The structure of the system

The structure of the system is presented here on the conceptual level, giving a user's view, without making any assumptions about the underlying implementation. The user's view is captured in a template for the arrangement of information about concepts and terms, information obtained from a search of multiple existing knowledge bases and/or from a common integrated knowledge base.

4.1 The SemWeb template

Figure 3 gives a rough outline of the SemWeb template, a list of frame slots (or slot groups) to organize information about concepts and terms. Some slots pertain primarily to terms as linguistic entities; others pertain primarily to concepts.

The template provides the basis for the common interface: The user starts with a template, fills in a term or concept (perhaps choosing from a classification displayed as a menu tree), and highlights the slots whose information she wishes to see. The system then selects and accesses the relevant sources, extracts the information needed, and presents the filled-in template to the user. The template helps the user identify the kind of information wanted and it provides the framework for integrating the information found and organizing it for display. The template also governs the system's internal workings: It provides the framework for organizing the system's knowledge about what information can be obtained from what knowledge bases and how to search each knowledge base. It serves as an input form for contributing data to the common distributed knowledge base. And finally, a frame hierarchy is one useful view of the internal structure of the common knowledge base.

The template focuses on information on individual concepts and terms. Many SemWeb functions also require views of overall conceptual / classificatory structures in various formats (linear listings, two-dimensional maps, etc.) with adequate browsing/

- Assist in the design and implementation of the **user interface, esp. choice of terms and icons**.

  Terms and icons must be chosen with the sometimes conflicting goals of communicating to the intended user group and of adhering to standards.
- Assist in the organization and formulation of **help messages and of documentation** and third-party software books.
- Serve as the **lexicon for machine translation** of interfaces and software-related documents
- **Assist the user in understanding interfaces and documentation, esp. in a foreign language**.
- **Support retrieval of software** for the end user or for **software reuse**.
- **Data element definition and standardization and organization of CASE tool databases.**
- All this functionality must be provided in **multiple languages** (for example, **software localization** for end users, **CASE tool databases for multinational development teams**)

Figure 1.
Semweb functions in a software context

navigation capability, allowing the user to move from a general overview to detailed classifications. The user needs access to multiple views, some showing the arrangement in present sources. Multiple views are essential to adapt to differing user requirements for conceptual arrangement, information given, main language, etc. Some views are grand structures of knowledge, such as the great library classification schemes; others are local overviews, such as the tables in the Longman Lexicon that represent the relationships between the various specialized terms for horse (filly, mare, stallion, etc) or the usage of various terms for father (dad, daddy, papa, etc.). An overview is referenced from its top concept. The **universal concept** serves as the reference point for universal classifications.

The system must include formal definitions that can capture fine nuances of meaning and usage. This is particularly important for establishing the proper correspondence between different languages and for relating similar functions in different software packages. Definitions of all meanings of a word can be arranged in a frame hierarchy (Chernyatin 1995).

- **Provide a semantic road map to individual fields and the relationships among fields; relate concepts to terms, and provide definitions**, clarify concepts by putting them in the context of a classification/ontology, relate concepts and terms across disciplines, languages, and cultures, thus providing orientation and serving as a reference tool.
- **Improve communication and learning generally**: Assist writers and readers, support learning through providing conceptual frameworks, support language learning and the development of instructional materials.
- **Provide the conceptual basis for the design of good research and implementation**. Assist researchers and practitioners in exploring the conceptual context of a research project, policy, plan, or implementation project and in **structuring the problem**. Consistent definition of variables and measures for more comparable and cumulative research and evaluation results.
- **Provide classification for action**: a classification of diseases for diagnosis, of medical procedures for insurance billing, of programmer skills for task assignments, of commodities for customs.
- **Support information retrieval**: **knowledge-based support of end-user searching** (menu trees, guided facet analysis of a search topic, browsing a hierarchy to identify search concepts, mapping from the user's query terms to descriptors used in one or more databases or to the multiple natural language expressions to be used for free-text searching), **hierarchically expanded searching**, support of **well-structured displays of search results**, providing a **tool for indexing** (vocabulary control, user-centered indexing).
- **Conceptual basis for knowledge-based systems**.
- **Do all this across multiple languages**
- **Mono-, bi-, or multilingual dictionary for human use**. **Dictionary/knowledge base for automated language processing** - machine translation and natural language understanding (data extraction, automatic abstracting/indexing).

Figure 2.
General SemWeb functions

Entry term, icon, concept, or group of terms, icons, or concepts (identified through a suitable identifier for the entity, preferably the system-wide identifier)

Other identifiers for the same entity

Broader and narrower frames (e.g., frame for a group to which the element belongs)

Spelling variants (other character strings in the same language)

Pronunciations (with dialect/regional variations and frequency information), in a phonetic alphabet or as digitized sound (to be used in a voice interface, for example).

Word root and derivation from the root

Compound terms, phrases, idioms of which the word is a part.

Etymological origin, history (leads to etymological cognates in other languages)

Part of speech, inflection rules, and other **syntactic information** (possible positions in a sentence, rules on combination with other terms to form expressions) (see below for semantically-based combination rules).

Terminological information: Other terms and icons with the same or similar meaning in the same language and in other (sub)languages/(sub)cultures/environments.

Definition and how-to description (as appropriate)

A preferred definition in English, French, etc., and other definitions and scope notes, given as text or hyperlinks to definitions found elsewhere. Definition in a **formal definition language**, possibly arranged in a frame hierarchy.

Semantic components, componential or feature analysis. Relevant feature space, necessary and sufficient features. Semantic root and derivation from the root.

For categories: Examples, prototype(s), members with degree of typicalness
For meanings that refer to concrete objects: a picture of the object and/or a picture that shows the designated object as part of a larger whole (as in a visual dictionary).

How-to descriptions (instructions how to execute the process or achieve the goal designated by the term, given as text or hyperlinks to definitions found elsewhere.

Usage notes, usage examples and quotations, familiarity and frequency information. For a group of terms that are close in meaning, subtle differences in meaning may be explained through examples and elucidation of definition and usage of the terms. Hyperlinks to text and program code in which the term occurs.

Disambiguation rules. Rules on how to determine the proper meaning of a homonym.

Category level (basic level, above basic level, below basic level).

Detailed conceptual relationships (Broader terms / hypernyms, narrower terms / hyponyms, parts / meronyms, whole / holonyms, concepts with which the concept at hand combines often / compound terms, etc.) and pointers to the concept's place in overall classificatory structures. Display of the structural relationships among subordinate concepts (a hierarchy, an association map, or a diagram showing relationships (relate to definition and usage, for examples see the Longman lexicon)

Rules on combination with other concepts to form expressions. For concepts that express relationships, especially verbs: A **case frame**. Slot filler restrictions in the case frame will define some aspects of usage.

Each piece of information, especially the concept or term itself and definitions, can be qualified by the (sub)language/(sub)culture/population group (including, for example, subcultures in the software domain) to which it applies and by the **audience level** for which it is suited.

Fig. 3. **Draft SemWeb template: Frame slots for information on concepts and terms**

EN drill n OED 1 (rivulet)	EN drill n W3 2			
EN drill n OED 2 (tool etc.)	EN drill n W3 5	... AHD 1	... RHD 1	FR drill n HD 2
EN drill n OED 3 (monkey)	EN drill n W3 6	... AHD 4	... RHD 4	FR drill n HD 1
EN drill n OED 4 (furrow)	EN drill n W3 7	... AHD 2	... RHD 2	
EN drill n OED 5 (fabric)	EN drill n W3 9	... AHD 3	... RHD 3	
EN drill v OED 1 (draw out)	EN drill v W3 9			
EN drill v OED 2 (trickle)	EN drill v W3 9			
EN drill v OED 3 (bore)	EN drill v W3 4	... AHD 1	... RHD 1	
EN drill v OED 4 (sow)	EN drill v W3 8	... AHD 2	... RHD 2	

Figure 4. **Word identifiers**: Language; character string; part of speech; source; word no.

(1) EN drill n OED 2.1 (tool)
(2) EN drill n OED 2.4 (military exercise)
(3) EN drill n OED 2.5 (a person who drills others)
(4) EN drill n OED 2.6 (rigorous training)
(5) EN drill n OED 3.1 (Mandrillus leucophaeus)
(6) EN drill n AHD 1.2 (disciplined, repetitious exercise, esp. military) (includes 2 and 4)
(7) EN drill n AHD 1.3 (specific exercise designed to develop a skill) (broader than 4?)
(8) EN drill n AHD 4.1 (Mandrillus leucophaeus)
(9) EN drill n RHD 1.2 (military exercise)
(10) EN drill n RHD 4.1 (Mandrillus leucophaeus)
(11) FR drill n HD 1.1 (Mandrillus leucophaeus)
(12) FR drill n HD 2.1 (military exercise)
(13) EN button n AHD 1.1 (disk-shaped fastener)
(14) EN button n AHD 1.3a1 (push-button switch)
(15) EN button n AHD 1.3a2 (interface element)

Figure 5. **Concept identifiers**: Terms with sense discriminators of the form .x

4.2 Entity types and entity identifiers

The SemWeb template gives a broad picture of the system's conceptual schema. We need to further specify the entity types for which frame instances can be created:

- **Character strings**
- **Terms** (words and phrases, including idioms and slang expressions)
  Linguistic roots and derived terms in both stem and inflected forms
- **Icons**
- **Concepts** (semantic roots and concepts derived through semantic modifiers)
- **Groups/classes of words/terms, icons, or concepts** for which some common assertions hold. Examples: All fifth declension Latin nouns; all English verbs that agree in their conjugated forms with sing (sing, ring, drink, etc.); all adjectives that could mean either a color or a race (such as white and black) and consequently share a semantic rule: They refer to a color when they qualify a non-human entity and to a race when they qualify a human entity. Frames for groups can represent grammatical knowledge in the same format as lexical knowledge.

A flexible system for identifying words/terms and concepts that uses the identifiers given in existing knowledge bases and is therefore compatible with the coexistence of many independent knowledge bases is shown in Figures 4 and 5. Note that the same character string may refer to different words in different sources. The word number distinguishes several words in one language represented by the same character string occurring in the same source. Terms or concepts can also be identified by a source ID and a

unique term or concept number within a source.

The identifiers thus constructed are unambiguous but not unique; a word or a concept has as many identifiers as it has sources. A common knowledge base will establish a correspondence between the different identifiers for the same word and likewise for the different identifiers for the same concept and, to the extent feasible, establish a system-wide identifier, which has the same form with the source ID for the system itself. Group entities require specially constructed identifiers.

4.3 Conceptual elements of the software domain

An ontological or lexical knowledge base needs a domain model which lists the entity types and possible relations between them. The following list gives, by way of example, some entity types in the domain of software.

function

external function

generalized function (e.g., block move)

application-specific function (e.g., spell check)

application

Values are functions defined on a general level, such as dealing with text (which could be part of DBMS)

software type (as defined by the focal application)

internal function (e.g., sort, string match)

user interface element

software element (e.g. object, data type)

hardware element (e.g. monitor, graphics card)

Specific hardware and software model (e.g. Apple Powerbook 503c, Word 7)

5 Sources of information for the proposed system

A tremendous amount of information has been amassed and codified in many existing sources. The system will provide access to as many of these as possible. A common knowledge base will — incrementally over time — include as much of this information as is feasible under legal restrictions and limitations of processing.

- **Lexicons and ontologies** from linguistic projects and knowledge-based systems.
- **Monolingual, bilingual, and multilingual dictionaries**, both general and specialized, including guides to usage (e.g. Fowler's) and guides to concepts (e.g. Kohl 1992).
- **Terminological standards**.
- **Order systems / subject access vocabularies** (thesauri, classification schemes / ontologies, etc.) used for information retrieval and other purposes.
- Tables of contents and indexes of books, such as software manuals and third-party software books.
- **Data dictionaries** of large information systems.
- **Object hierarchies** of software systems, where objects can represent data types, any kind of domain objects, or software functions.
- **Laws and regulations** (food regulations contain definitions for many foods, drug laws classify drugs into "schedules" based on their psychoactive effects, etc.)

6 Development of the SemWeb software and of common SemWeb knowledge bases

The development of SemWeb requires incremental work on a number of major tasks. Fortunately, development can build on many projects already underway; SemWeb would bring their results together in a unified framework. Two principles make a system of this magnitude possible: **multiple contributors** and **virtual integration**, the principles on which the World Wide Web itself and systems like OCLC and software like LINUX are based.

Tasks required for the development of the software for implementing the common interface

- **Develop a "super standard" for any kind of ontological and lexical information (information on concepts and terms**). A good template/frame structure is central to the success of the system. While there exist standard formats for machine-readable dictionaries, subject authority files, and classification data, there is no one format for **all** the types of data on concepts and terms as envisioned here. The existing standards must be brought together and augmented to accommodate even very specialized lexical projects.
- **Develop a conceptual schema for a database of ontological and lexical knowledge bases**, whether machine-readable or printed. Each SemWeb user needs access to such a database, which would record such things as the domain, the kind of objects (terms, icons, concepts) covered, the kind of information given, address and access protocol, copyright and access fees).
- **Develop the software** for the system: a **kernel** (gets the user's request, selects the knowledge bases to be searched, integrates the information found, and displays it to the user) and **special modules for searching specific knowledge bases**. The master list serves as a knowledge base for this software, including information on user fees and on copyright status (to determine whether information can be copied into the common knowledge base or can be included only by reference).

Tasks for the incremental development of a common integrated knowledge base

- Develop a **master list of ontological and lexical knowledge bases** for the domain of the common integrated SemWeb knowledge base. This would be based on existing partial lists. This master list could reside in a central place or on each user's computer. Each user (or group of users could have a complementary database of additional resources (esp. private resources not accessible outside the group)

  The development of a common knowledge base can proceed on three levels

- Level 1: **The system keeps the concept or term records it assembles** in response to a user request, replacing copyrighted information with a reference, and keeps a directory of these records, regardless of where they are stored. For the next request for the same concept or term, only the knowledge bases that have been updated or that contain copyrighted information need to be accessed again.
- Level 2: **The user can edit a concept or term record and store the edited copy, thus adding value** by establishing correspondences between the numbered entries from several dictionaries (see Section 4.2) and/or creating system-wide identifiers for words and word senses; establishing correspondences with concepts in classification schemes; synthesizing a natural language definition that is better than any of the definitions found in dictionaries; creating a frame hierarchy of formal definitions of word senses. Some of this editing must be done anyway before the information gathered from several sources can be used. The system allows users to share the fruits of their labor.
- Level 3: **Development of a well-structured knowledge base that integrates knowledge from many sources**. The structure of such a knowledge base must be designed in accordance with the super standard mentioned above. The integration of information from various sources can be automated to a large extent, using intelligent software that builds on existing work. The information produced through editing would also be used, with the structure of the Web pages facilitating such incorporation. The system could acquire further knowledge from the analysis of text and of term use in searching.

SemWeb is conceived as a federated system with multiple collaborators by subject, application, or language specialization, and with data distributed over multiple sites but appearing to the user as a unified system. Each contributor and each user has a status (which might include description of expertise along several dimensions); some collaborators might be recognized as official contributors, others might just use the knowledge base for their project - the system would provide an environment for the more efficient development of specialized concept and terminology knowledge bases, while at the same time providing efficient storage and wide access to the results. More casual users could also add their own information and suggest additions and corrections, with mechanisms for quality control. Users could restrict retrieval to information entered or reviewed by a contributor meeting certain status requirements.

7 Access to SemWeb

Access to SemWeb would be provided in multiple ways:

- Through the World Wide Web, intranets, and other online means.
- Through integration into search systems, giving transparent ("behind the scenes") assistance or explicit system suggestions to be modified as needed.
- Through products derived from it (special classifications, dictionaries, etc.).

8 Conclusion

By providing integrated access to a wide variety of ontological and lexical knowledge bases and by providing a forum in which users can augment these resources by sharing ontological/classificatory and lexical knowledge, a SemWeb system creates the conceptual infrastructure that is urgently needed to reap the full benefit from the information exchange made possible by today's information infrastructure.

Bibliography

Sample dictionaries and other lexical resources

AHD The American Heritage dictionary of the English language.

HD Harraps new college French and English Dictionary

OED The Oxford English dictionary.

RHD The Random House dictionary of the English language.

W3 Webster's third new international dictionary.

Fowler, H. W.; Gowes, Ernest Sir, revisor. **A dictionary of modern English Usage**. New York and Oxford: Oxford University Press; 1965.

McArthur, Tom . **Longman lexicon of contemporary English**.

Kohl, Herbert (1992). **From archetype to Zeitgeist. Powerful ideas for powerful thinking**. Boston: Little, Brown, and Company.

WordNet.
Home page:
http://www.cogsci.princeton.edu
Good search:
http://www.notredame.ac.jp/cgi-bib/wn.cgi

Upper Cyc Ontology
http://www.cyc.com/cyc-2-1/cover.html

Unified Medical Language System (UMLS).
http://www.etb.nlm.nih.gov//sampler/umls.html

Kirby, D. G.; Borgeest, M. **US government dictionaries: a selective guide**. Reference Services Review 22(3), 33-68.

Margolis, P. (1996) **Random House personal computer dictionary**. New York, NY: Random House; 1995. 528 p. ISBN 0-679-76424-0

Computing and information technology French dictionary. Teddington, Middlesex: Peter Collin; 1996. 494 p.

English-Japanese dictionary of computer. Tokyo:

Nichigai Asoshietsu; 1996. 1173 p.

Ferreti, Vittorio (1996). **Dictionary of computing. English-German, German-English.** Berlin, New York: Springer; 1996. 1370 p.

Standards for lexical knowledge bases

ISO CD 12620.2: **Computational aids in terminology — Data categories.**

ISO DIS 12200, **Computational aids in terminology — Terminology Interchange Format (TIF) — An SGML application.**

USMARC format for authority data: including guidelines for content designation. / Prepared by the Network Development and MARC Standards Office. Washington, DC: Library of Congress. Cataloging Distribution Service; 1993.

USMARC format for classification data: including guidelines for content designation. / Prepared by the Network Development and MARC Standards Office. Washington, DC: Library of Congress. Cataloging Distribution Service; 1991.

Other references

Chernyatin, Valentin; Zimmermann, Harold (1995). **Towards a multilingual semantic database**. Draft (email: gg15hzhz@rz.uni-sb.de).

Conlon, S. P. N.; Evens, M.; Ahlswede, T. (1993). **Developing a large lexical database for information retrieval, parsing, and text generation systems**. Information Processing & Management, 29(4), 415-31.

Karkaletsis, V.; Spyropoulos, C. D.; Vouros, G.; Halatsis, C. (1995). **Organization and exploitation of terminological knowledge in software localization**. TermNet News - Journal of the International Network for Terminology, no. 42, 42-48.

Pappano, Laura (1996). **Publisher plans to book 'Library of Language' on the Internet**. [On plans by Merriam-Webster to make the Webster's Fourth database available on the Internet and solicit comments from the public.] Washington Post, 1996 April 28, A3.

Senez, Dorothy (1995). **Developments in Systran**. [Reports on the joint use of Systran's own lexicon and Eurodicautom, the European Communities multilingual terminology database, in machine translation.] Aslib Proceedings, 47(3), 99-107.

Strehlow, Richard A., and Wright, Sue Ellen, eds. (1993). **Standardizing terminology for better communication: practice, applied theory, and results.** Philadelphia: ASTM.
Especially
Wright, Sue Ellen, and Melby, Alan K., TEI-TERM: A proposed format for the interchange of terminology data using standard generalized markup language. 200-216.

Wright, Sue Ellen, and Strehlow, Richard A., eds. (1995). Standardizing and harmonizing terminology: Theory and practice. Philadelphia: ASTM.
Especially
Galinski, Christian. Exchange of standardized terminologies within the framework of the Standardized Terminology Exchange Network (STEN). 141-154.

Wright, Sue Ellen. Creating a data element dictionary for computer-aided terminology work. 169-186.

Melby, Alan K. Implementing the Terminology Interchange Format. 187-199.