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 or soergel@umail.umd.edu

1996-11-14

SemWeb

Proposal for an

Open, multifunctional, multilingual, system for integrated access to knowledge about concepts and terminology

Exploration and development of the concept

Expanded version of a paper published in Advances in Knowledge Organization v.5 (1996): 165-173 (4th Annual ISKO Conference, Washington, D.C., 1996 July 15-18). This version has more detail, particularly in Figure 1 and Section 4.1.2.

Abstract. This paper presents a proposal for the long-range development of an open, multifunctional, multilingual system for integrated access to many kinds of knowledge about concepts and terminology. The system would draw on existing knowledge bases that are accessible through the Internet or on CD-ROM and on a common integrated distributed knowledge base that would grow incrementally over time. Existing knowledge bases would be accessed through a common interface that would search several knowledge bases, collate the data into a common format, and present them to the user. The common integrated distributed knowledge base would provide an environment in which many contributors could carry out classification and terminological projects more efficiently, with the results available in a common format. Over time, data from other knowledge bases could be incorporated into the common knowledge base, either by actual transfer (provided the knowledge base producers are willing) or by reference through a link. Either way, such incorporation requires intellectual work but allows for tighter integration than common interface access to multiple knowledge bases. Each piece of information in the common knowledge base will have all its sources attached, providing an acknowledgment mechanism that gives due credit to all contributors. The whole system would be designed to be usable by many levels of users for improved information exchange.

1 Introduction

"The global information society of the 21st century will rely increasingly on an information infrastructure which must have two essential components: the global telecommunication and electronic networks epitomized by the Internet; and, underpinning it, a conceptual infrastructure reflecting the way knowledge and information are organized." (From the recommendations of the ISKO/Polish Information Society Joint Seminar on the Compatibility of Order Systems, Warsaw, Poland, Sept. 13-15, 1995)

This paper deals with the intellectual infrastructure. It presents the vision and basic architecture of a system for integrated access to data on concepts and terminology. The system would bring together for the user data from a variety of sources that up to now exist largely in separate worlds, including dictionaries, thesauri, and classification schemes; it would draw on existing knowledge bases that are accessible through the Internet or on CD-ROM and on a common integrated distributed knowledge base that would grow incrementally over time. Existing knowledge bases would be accessed through a common interface that would search several knowledge bases, collate the data into a common format, and present them to the user. The common integrated distributed knowledge base would provide an environment in which many contributors could carry out classification and terminological projects more efficiently, with the results available in a common format. Over time, data from other knowledge bases could be incorporated into the common knowledge base, either by actual transfer (provided the knowledge base producers are willing) or by reference through a link. Either way, such incorporation requires intellectual work but allows for tighter integration than common interface access to multiple knowledge bases. Each piece of information in the common knowledge base will have all its sources attached, providing an acknowledgment mechanism that gives due credit to all contributors. The whole system would be designed to be usable by many levels of users for improved information exchange.

Implementation of such a system is clearly a long-range proposition that would require a collective effort by a number of people. People interested in participating in such a project are invited to communicate with the author.

2 Rationale for the proposed access system and knowledge base

There are now a multiplicity of order systems and classifications, terminological knowledge bases, and linguistic dictionaries, each serving a limited purpose but widely overlapping in their content. The proposed common interface would provide simultaneous access to all concept and terminology knowledge bases on the Internet with limited integration of information "on the fly"; going a step further, the proposed integrated distributed knowledge base would provide a home for all of these knowledge bases who care to join, eliminating duplication while preserving the integrity of each source and establishing relationships across sources; . As outlined below, the proposed 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, the information to be added for each additional function becomes less and less for each function added, and it is precisely the multifunctionality that makes the considerable investment pay off. Since this is an investment into infrastructure, it should properly come from the public rather than the private sector. Thus, the proposed knowledge base 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 knowledge bases of concepts and terminology 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 the environment provided by the proposed knowledge base.

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

3 Functions of a knowledge base on concepts and terminology

Knowledge on concepts and terminology, especially knowledge on classificatory structure, can serve one or more of the functions listed in Figure 1. The proposed system could serve all of these functions directly. It could also serve as an environment for projects developing order systems (thesauri/classifications), concordances between order systems, linguistic dictionaries, etc.; in the ideal case, specialized ordering systems, dictionaries, etc., both machine-readable and printed, could be produced simply by extraction from the integrated knowledge base.

- The **basic functions** of a knowledge base on concepts and terminology — on which all others rest — are **to map out a concept space, to relate concepts to terms, and to provide definitions**, thus providing orientation and serving as a reference tool.
    - To provide a *semantic road map and common language* for an individual field and, perhaps more importantly, map the relationships among fields.
    - To *clarify concepts by putting them in the context of a classification*/ typology and to provide a system of definitions.
    - To *relate concepts and terms across disciplines, languages, and cultures.*

    Many specific functions build on this foundation.
- **To improve communication and learning generally**.
    - To *assist writers* in conceptualizing a topic and in finding the proper term.
    - To *assist readers* in understanding text.
    - To *support learning* about any topic *by providing* the student with a *coherent conceptual framework* appropriate to the students age and background. Such learning may occur in the course of a search for information where the conceptual framework provided helps the student in defining a good approach to searching.
    - To *support language learning*, esp. learning of a foreign language.
    - To *assist in the preparation of instructional materials* by providing didactically useful arrangements of topics.

Figure 1. **Functions of a knowledge base on concepts and terminology**

- **To assist researchers and practitioners in exploring the conceptual context of a research or practical problem** — a research project, policy, plan, or implementation project — and in **structuring the problem**, thereby providing the conceptual basis for the design of good research and implementation.
    - To *present the issues in a field or application area in a coherent framework.*
    - To provide classification and *consistent definition of variables for research* — would support comparison of research results and make research more cumulative.
    - To *assist in problem-solving*: Assist in the exploration of the dimensions of a problem and aspects to be considered in its solution, to provide a classification of approaches to solving problems (for example, a classification of approaches to drug abuse prevention as a help in designing drug abuse prevention projects), and to assist in the definition of consistent, evaluation criteria, thus enhancing comparability of evaluations.
- **To provide classification for action, for example**
    - Disease classification to assist with diagnosis.
    - Classification of medical procedures for insurance purposes.
    - Classification of medical outcomes to assist with treatment evaluation.
    - Customs classification to determine the correct import or export duty.

Figure 1. **Functions of a knowledge base on concepts and terminology**, cont.

- **To support information retrieval**
  - To provide a **tool for searching** in printed indexes or in computer search systems, particularly knowledge-based support for end-user searching on the Internet and other online services. The knowledge base would support searching in multiple languages independent of the language used in each database; free-text searching; searching multiple databases using different index languages; searching in any kind of database — bibliographic, full-text and hypermedia, directory, numeric, etc. In particular, the knowledge base would support
    - the *elicitation of user needs*, through a series of menus based on search tree, or through *guidance in the conceptual analysis of a search topic* (questions based on a facet structure, presentation of a segment of the concept hierarchy);
    - *browsing the classification structure* to identify useful concepts for a search and to assist the user in identifying terms at the level of specificity desired (The user may not have command of the vocabulary needed.);
    - *mapping from the user's query terms to descriptors* used in a database *or to the multiple natural language expressions* to be used for free-text searching;
    - *inclusive* (hierarchically expanded) *searching*;
    - *enhanced ranking algorithms* that use information on concept and term relationships;
    - *searching multiple databases* by mapping the users query terms to the descriptors used in each of the databases, or mapping the descriptors used in querying one database to the descriptors to be used in querying the other databases (switching); common search language.
  - To **support information display**, especially presentation of search results:
    - *meaningful arrangement of units* (document records, paragraphs, property data on a given substance assembled from several databases), including knowledge-based clustering of records retrieved;
    - *meaningful arrangement of information within a record*.

Figure 1. **Functions of a knowledge base on concepts and terminology**, cont.

- **To support information retrieval**, continued
  - To **provide a tool for indexing**. In particular, the knowledge base would support
    - *vocabulary control*;
    - request-oriented (problem-oriented, *user-centered) indexing*;
    - indexing *several databases* in a field with a *common index language* and sharing the results of indexing to reduce overall indexing effort;
    - *mapping indexing descriptors from one system to another*.
  - To **facilitate the combination of multiple databases or unified access to multiple databases** through providing a *common search language* to multiple databases, providing a *common index language* for a number of databases in a field, or mapping indexing descriptors from one system to another.
- **To serve as the conceptual basis for knowledge-based systems.**
- **To serve as a dictionary - in monolingual, bilingual, and multilingual mode - for human use.**
- **To serve as the dictionary/knowledge base for automated language processing** - including machine translation, data extraction, automatic abstracting and indexing, and natural language understanding generally. It should be noted that parsing natural language requires not only morphological information and information about the possible syntactic roles of a term but also a great deal of semantic information.
  - Special case: Spell check dictionary, knowledge base for grammar checking.
- **To provide a classification/ontology for data element standardization.**

Figure 1. **Functions of a knowledge base on concepts and terminology**, cont.

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. Some elements in the template are more suitable for use with the tighter integration and structuring of information in a common integrated knowledge base.

4.1 The SemWeb template

Figure 2 gives the first draft of an overall outline of the SemWeb template or frame, a list of frame slots (or slot groups) to organize information about concepts and terms. A frame instance can be established for any value of any of the entity types listed in Section 4.2. Some slots pertain exclusively or primarily to terms as linguistic entities; others pertain exclusively or primarily to concepts. Many of the slot fillers will be references to other frames. (The information dealt with in SemWeb could also be specified on an "atomic" level by giving the entity types and relationship types used in the system to form assertions. However, our purpose here is better served by the holistic template or frame view.)

The entity types discussed in Section 4.2 include groups of terms or concepts, such as 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. A frame for a groups represents grammatical knowledge and knowledge about generalized conceptual relationships (including generalized case frames for groups of verbs. Once it is fully worked out, **the SemWeb template thus provides a unifying mechanism for representing both lexical and grammatical knowledge**.

The template provides the basis for the common interface: The user starts with a template, fills in a term or concept identifier (possibly choosing from a classification displayed as a menu tree), and highlights the slots whose information she wishes to see. The system then accesses all relevant sources it knows about, 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 provides the basis for 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.

A good template or frame structure is central to the success of the proposed system. While there exist standard formats for machine-readable dictionaries and subject authority files (see the bibliography), there is no one format that structures **all** the types of data on concepts and terms as envisioned here.

The template focuses on information on individual concepts and terms. To support many of the functions listed above, the system must also show overall conceptual / classificatory structures in various formats (linear listings, two-dimensional maps, etc.) with adequate browsing/navigation capability allowing the user to move from a general overview to detailed classifications. The system should provide access to multiple views, some corresponding to the arrangement in present sources. Multiple views are essential to make an integrated system workable for multiple groups of users, each with different requirements for conceptual arrangement, information given, main language, etc. Some of these 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 a number of terms, such as 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 in the appropriate slot for the most specific concept that still covers its entire scope. A frame for the *universal concept* serves as the reference point for universal classifications.

The system must provide for formal definitions that can capture fine nuances of meaning and usage. This is particularly important for establishing the proper correspondence between different languages. Definitions of the various meanings of a word might be arranged in a frame hierarchy as proposed in Chernyatin (1995).

Entry term, concept, or group of terms 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

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.

Word root and derivation from the root

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

Etymological origin, history (from this etymological cognates in other languages can be inferred and displayed)

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

Terminological information: Other terms with the same or similar meaning in the same language and in other languages.

Definitions

> Verbal definitions and scope notes in many languages from many sources. In an integrated system: A preferred definition in each language.
>
> Definitions 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).

Figure 2. **The SemWeb template: Frame slots for information on concepts and terms**

Usage

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 text elaborating on the definition and usage of the terms, with examples.

Category level (basic level, above basic level, below basic level), qualified by population and population subgroup

Detailed conceptual relationships (Broader terms / hypernyms, narrower terms / hyponyms, parts / meronyms, the whole of which the concept is a part / 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 or table showing relationships (closely linked with 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.

Figure 2. **The SemWeb template: Frame slots for information on concepts and terms**, continued

4.1.1 **Definitions**

Formal definition(s) would be given in a special definition language (possibly restricted English). The formal definition is intended to allow automatic derivation of concept relationships. The formal definitions may take the form of frames, one frame for each meaning; commonalities among several meanings will be expressed through a frame hierarchy. Figure 2a shows an example.

The highest level definition for *EN drill n OED 2 (tool etc.)* could read something like this

> Repetitive, usually somewhat laborious, action, possibly using an instrument, with a goal, or the tool used for such action, or a person implementing the action, or a person supervising the action. The action may be further classified by context.

This sets up a frame with five slots. The slot that gives rise to isa will be identified with a *

Action	:
Tool or implementer	:
Supervisor of action	:
Goal	:
Context:	:

EN drill n OED 2.1 (tool)		
Action	:	Turning
Tool or implementer*	:	Rod or other sharp object or a machine turning the object
Goal	:	A hole in some material
Context	:	Not specified

EN drill n OED 2.4 (military exercise)		
Action*	:	Repetitive motion
Tool or implementer	:	A soldier or a group of soldiers
Goal	:	Learning certain movements or actions so that they become automatic
Context:	:	Military

EN drill n OED 2.5 (a person who drills others)		
Action	:	Repetitive motion or repetitive execution of mental task Default: Repetitive motion (90 %)
Tool or implementer	:	A person or group of persons
Supervisor of action*	:	A person
Goal	:	Learning certain movements or actions so that they become automatic
Context	:	Default: Military (90 %)

EN drill n OED 2.6 (rigorous training)		
Action	:	Repetitive motion or repetitive execution of mental task
Tool or implementer	:	Person or group of persons
Goal	:	Learning certain movements or mental tasks so that they become automatic or can be easily done at will
Context	:	Not specified

Figure 2a. **A frame hierarchy of formal definitions**

4.2 Entity types and entity identifiers

The conceptual structure of a domain can be captured most accurately by a conceptual schema identifying the entity types and relationship types covered. The SemWeb template gives a broad picture of the system's conceptual schema, and this paper does not enter into a detailed discussion of relationship types beyond that. But we must discuss the core entity types and the method for assigning entity identifiers, a subject that is somewhat dry and technical but vital to the smooth functioning of a system that relies on access to a number of independent knowledge bases.

The following is a list of entity types for which frame instances can be created:

- *Character strings*
- *Terms* (words and phrases, including idioms and slang expressions)
    - Linguistic roots
    - Terms derived terms from linguistic roots - stem form
    - Inflected forms of terms
- *Concepts*
    - Semantic roots
    - Concepts derived through semantic modifiers (Soergel 1991)
- *Groups/classes of words/terms 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 (including sing, ring, drink, sink, stink, swim, begin)

    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 3 and 4.

Figure 3 shows word identifiers consisting of

Language indicator; character string; part of speech; source; word number

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

Figure 3. **Word identifiers**: Language; character string; part of speech; source; word no.

The source is needed to completely identify a word since the same character string may refer to different words in different sources. The word number serves to distinguish several words in one language represented by the same character string occurring in the same source. In addition, words or terms can be identified through a unique word or term number within a knowledge base; for system-wide use, such internal term numbers must be prefixed by the code for the source.

Figure 4 shows concept identifiers consisting of

> Language indicator; character string; part of speech; source; word + sense number, optional sense discriminator

Alternatively, a concept can be identified through a special notation, as in a classification schemes, or through a concept number specific to a given knowledge base; for system-wide use, either must be prefixed by the code of the source.

(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)

Figure 4. **Concept identifiers**: Terms with sense discriminators of the form .x

The word and concept identifiers thus constructed are unambiguous but not unique; a word or a concept has as many identifiers as it has sources. The 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.

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. The 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, etc.) used for information retrieval and other purposes.
- *Data dictionaries* of large information systems.
- *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 SemWeb

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 common interface

- *Develop a "super standard" for any kind of 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 comprehensive master list* of machine-readable and printed resources, each described following a standard schema, based on existing partial lists. This master list would keep track of all the update schedules.

- *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). The user would be informed before the system follows a link that would incur charges.

Tasks for the incremental development of a common integrated knowledge base

- 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*. Such editing might include any of the following:

    - Establishing correspondences between the numbered entries from several dictionaries (see Section 4.1) 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 form several sources can be used. The system allows users to share the fruit of their labors.

- 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 and other online means.
- Through integration into search systems, giving — at the user's option — transparent ("behind the scenes") assistance or explicit system suggestions to be modified as needed.
- Multiple views, some corresponding to the arrangement in present sources. Multiple views are essential to make an integrated system workable for multiple groups of users, each with different requirements for conceptual arrangement, information given, main language, etc.
- Through products derived from it (special classifications, dictionaries, etc.).

8 Conclusion

By providing integrated access to a wide variety of lexical and classification knowledge bases and by providing a forum in which users can augment these resources by sharing lexical and classificatory knowledge, SemWeb creates the conceptual infrastructure that is urgently needed to reap the full benefit from global information exchange that is driven by the same information infrastructure that makes SemWeb possible.

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.

Tom McArthur. **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

Unified Medical Language System (UMLS) (brings together and relates a number of lexical resources in biomedicine).
http://wwwetb.nlm.nih.gov//sampler/umls.html

Dewey Decimal Classification

Library of Congress Classification

Library of Congress Subject Headings

Dictionary of Occupational Titles

International Classification of Diseases

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

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.

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.

Soergel, Dagobert (1991). **Beyond facets: Semantic roots and modifiers.** Proceedings of the 2nd American Society for Information Science/SIG-CR Classification Research Workshop. Held at the 54th ASIS Annual Meeting, Washington, DC, Oct. 27, 1991. Washington, DC: ASIS. p.149-158. (Advances in Classification Research. v. 3)

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.