SemWeb: Integrated access to distributed ontological resources

1 Introduction and overview

We propose to develop a system, dubbed SemWeb (semantic web), that would revolutionize the way people — from experts to students — interact with conceptual structures and terminology and the way they share such knowledge. Assume you are planning a campaign for preventing alcohol abuse and are interested in the factors influencing drinking behavior — from neurochemical to social and legal. Wouldn’t it be helpful to be able to find a diagram of the causal relationships among these factors (a form of concept map) that summarizes research from many disciplines. Wouldn’t it be more helpful still to be able to click on a chemistry term you do not quite understand and get a definition and a concept hierarchy around the term or, to pursue this more deeply, side-by-side definitions and hierarchies from several sources. Repeat that for a legal term — SemWeb would access different sources to get information about it. If you wanted to search a digital library, you could request a complete list of synonyms, collated from several sources. If you wanted to search a bibliographic database, you could request the descriptor from its language as a starting point for your query. Once you have refined your query for this database, you could request that SemWeb translate it into the vocabulary of another database.

Or picture a linguist who needs to put together a machine-readable dictionary for summarizing texts in a very specialized domain in several languages. SemWeb could access several specialized vocabularies, including bilingual glossaries, to compile a multilingual sublanguage dictionary (with translations appropriate to the domain) and then get syntactic information needed for natural language processing (NLP) from large machine-readable dictionaries in each language. Manual editing would still be needed, but the effort needed to produce a high-quality dictionary would be significantly reduced. In the SemWeb framework, the fruits of this labor would be available to other linguists in need of a dictionary in this or a related domain. In fact, the dictionary would be useful for cross-language searching of any database as long as the query topic is in the domain. If the linguist created hierarchical relationships among concepts in the domain, these might be useful to an instructor in a related course for structuring course materials.

Or picture a teacher who needs to communicate complex interrelationships to his students. He could find concept maps and present them on paper or, even better, online to afford the students the same interactivity described above. Students could integrate their knowledge by drawing their own concept maps, either individually or collaboratively. The teacher may also test students’ understanding of some topic by having them draw concept maps. SemWeb could score the quality of these concept maps (or at least assist in the scoring), drawing on a set of evaluation criteria and on conceptual and terminological knowledge from various sources.

The SemWeb proposal is based on the synergistic exploitation of existing ontologies, classification schemes, lexica, thesauri and dictionaries and their vast intellectual capital, which is now largely untapped for lack of access. This exploitation requires knowledge integration. We envision two approaches working together: (1) automating knowledge integration to the extent feasible and (2) providing tools to facilitate knowledge integration by users for their own purposes, but in a format that can be shared with others. The architecture of the proposed SemWeb access component is shown in Figure 1 and more fully explained in Section 5.4.
We expect that such a system would have a significant impact on how people think and structure problems; how they retrieve and organize information; how they communicate and collaborate across disciplines, languages, and cultures; and how they learn. By making vast amounts of knowledge accessible, it would also expand the horizons on what can be done in knowledge-based systems and natural language processing. All of this contributes to scientific productivity.

We propose to build such a system, a software package that could be deployed in many contexts, and to install a demonstration prototype Web server. We also propose to study basic issues that arise in the system’s design, as well as the use and impact of the prototype. All of this is a challenge. We believe that we have assembled an interdisciplinary team that can meet this challenge, bringing together ideas and approaches from information science, linguistics, AI, and education. We are asking for the resources that would enable us to develop the basic SemWeb idea into a design, to implement the design, and to do preliminary studies of use and impact. Following the underlying principle of SemWeb, we will build on whatever is already available. We will also try to further enhance the results by communicating with others and encouraging them to build on and extend what we can do. We already have an expression of interest from GMD/IPS in Germany (see attached letter).
2 Objectives

Overall objective: Develop and study a knowledge network that supports working with concepts and their relationships; facilitates knowledge sharing and integration across disciplines, languages, and cultures; and serves to increase scientific productivity and to improve learning.

Specific objectives:

1. Design and develop a system for (1) integrated access to a wide variety of distributed lexical and ontological/classificatory resources — bridging disciplines, languages, cultures, and functions — for a wide variety of uses and users (including education, information retrieval (IR), knowledge-based systems and natural language processing (NLP)) in a networked environment and (2) collaborative development and maintenance of ontologies and lexica.

2. Improve methods of visualizing concept relationships for smooth information transfer.

3. Define measures and methods for the evaluation of ontologies, lexica, and their representations and for correlating and integrating ontologies.

4. Implement a prototype system accessible through the World Wide Web.

5. Study the use and impact of the prototype through pilot applications and user studies in NLP, IR, and education (especially the effect of concept maps on learning).

6. Test collaborative methods involving several communities in the system’s development.

We now pose the problem (Section 3), describe the perspectives the principal investigators bring to a solution (Sec. 4), lay out the proposed SemWeb architecture (Sec. 5), detail the R&D tasks (Sec. 6), and look ahead at some projects that could build on and expand SemWeb (Sec. 7).

3 Background: The problem

3.1 Importance of lexical and ontological knowledge bases

Ontological and lexical structures are the underpinning of scientific and scholarly work, of learning, and of machine intelligence. They serve many critical functions in thinking and in communicating, organizing and retrieving information by people and machines. The functions of tools providing such structures (dictionaries, thesauri, ontologies/classifications) include the following:

- 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 or icons across disciplines, languages, and cultures; thus providing orientation and serving as a reference tool.

- Improve communication and learning: Assist writers and readers, support learning through providing conceptual frameworks and challenging students to produce such frameworks, support language learning, and support 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*. Support consistent definition of variables/measures for more comparable and *cumulative research* and evaluation results.

• **Provide classification for action:** a classification of diseases for diagnosis, of medical procedures for billing, of staff skills for task assignments, of commodities for customs.

• **Support information retrieval:** provide knowledge-based support of end-user searching (menu trees, guided facet analysis of a search topic, browsing a hierarchy or concept map 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 for free-text searching); support *hierarchically expanded searching*; support *well-structured displays of search results*; provide a **tool for indexing** (vocabulary control, user-centered or problem-oriented indexing).

• **Provide a conceptual basis for knowledge-based systems.**

• **Do all this across disciplines, languages, and cultures.**

• **Serve as mono-, bi-, or multilingual dictionary for human use** and as dictionary/knowledge base for natural language processing — machine translation and natural language understanding for data extraction and automatic abstracting/indexing.

This analysis leads to the working hypothesis of this proposal, that improvements in the structure and accessibility of such knowledge bases will help people organize their thoughts in research, problem solving, and decision making and have a significant impact on information sharing, retrieval, and access and on information acquisition, analysis, evaluation, and presentation. It will ultimately improve learning and scientific productivity. An integrated access system across domains such as SemWeb will foster the linking of ideas across domains.

### 3.2 The present status of lexical and ontological knowledge bases

The present status of lexical and ontological knowledge bases can best be characterized as splintered. They come in many forms (dictionaries and lexica; glossaries and terminological standards, some found in laws and regulations; thesauri; ontologies, classification schemes, taxonomies, and typologies; concept maps; data dictionaries and object hierarchies of large information systems and software systems). They are usually developed for a specific, limited purpose and restricted in their use. But they overlap widely in their content: knowledge bases developed for one function could often be reused for other functions, if only they were complemented with additional information (often available in other sources) and then reformatted for the purpose at hand. A further obstacle to the exploitation of these resources is limited accessibility and lack of knowledge about resources available. As a result, the intellectual capital accumulated in these resources is underused. Furthermore, new lexical and ontological knowledge bases are often developed without taking full advantage of what already exists, thus expending more effort than necessary, yet ending up with a result that is potentially incompatible with existing schemes. In sum, the support for the critical functions outlined in Section 3.1 is far from what it could be. We want to change that.
3.3 Existing approaches to a solution

Most relevant are projects that access multiple sources and/or support collaborative development.

The Stanford Ontolingua server (http://ontolingua.stanford.edu/, Farquhar et al., 1996) provides a Web-based environment for the construction of ontologies which are organized as frame hierarchies with a rich semantics. It also supports browsing and comparing ontologies on the server (currently about 60, mostly small). One can find all ontology entries for a term, with hyperlinks. In building an ontology, one can incorporate existing ontologies (or pieces).

The Unified Medical Language System (www.nlm.nih.gov/pubs/factsheets/umlskss.html, NLM 1998) is a thesaurus database that integrates more than 40 source vocabularies from the medical domain in its Metathesaurus. It also includes a Specialist Lexicon with syntactic information for NLP. The Semantic Network provides conceptual organization. UMLS can be queried through a Web-based interface and returns information from several sources. It is also used for search assistance in Internet Grateful Med (www.nlm.nih.gov/pubs/factsheets/igm.html).

The Language Representation Database Project (www.asel.udel.edu/natlang/nlp/lrd.html, Zickus 1995) creates a central database of lexical sources (e.g., WordNet, the Brown Corpus, a phonetic module) from which customized lexicons can be extracted.

One-look dictionaries. The Faster Finder (www.onelook.com) is a server that searches 272 dictionaries (with a total of 1,720,143 words); it returns a list of dictionaries that cover the search word, with hyperlinks to the dictionary entries. Wordbot can query dictionaries and accepts user additions or comments (www.cs.washington.edu/homes/kgolden/dictionaries.html).

Many projects address the problem by constructing widely usable large lexica or ontologies. Examples are WordNet (www.cogsci.princeton.edu/~wn), a widely available lexical database/thesaurus incorporating a concept hierarchy used in many NLP and IR projects; EuroWordNet (www.let.uva.nl/~ewn), a multilingual database with basic semantic relations between words for Dutch, Italian and Spanish linked to WordNet; the EDR Japanese and English Electronic Dictionary, which includes a concept dictionary; COMLEX (for Common Lexicon), a project of LDC (see below) that creates a set of lexical databases focused on particular functions, licensing a morpho-syntactic and a pronunciation lexicon (PRONLEX); ACQUILEX (Calzolari, et al 1995); the CYC Ontology (www.cyc.comm); SENSUS (Hovy); and the very widely used large classification schemes Dewey Decimal Classification (DDC, www.oclc.org/fp for info) and the Library of Congress Classification (LCC), both available on CD-ROM.

There are many efforts at coordination and referral to resources. Among them: The Linguistic Data Consortium (LDC, www.ldc.upenn.edu/Ldc) “creates, collects and distributes speech and text databases, lexicons, and other resources for research and development purposes.” The Association for Computational Linguistics’ ACL NLP/CL Universe (www.cs.columbia.edu/~radev/u/db/acl/html) is a referral site. The European Language Research Association (ELRA, www.icp.grenet.fr/ELRA/home.html) “promotes the creation, verification, and distribution of language resources in Europe”. TELRI (www.ids-mannheim.de/telri) works toward “creating a
viable infrastructure between leading European language and language technology centres”; so does the *European Network in Language and Speech* (ELSNET, www.elsnet.org). The *POINTER* final report (www.mcs.surrey.ac.uk/AI/pointer) outlines proposals to facilitate the distribution of terminologies (through ELRA), as well as their re-use in different contexts and for different purposes. The Working Group: Shared Vocabularies for Representing Pedagogical Knowledge (http://advlearn.lrdc.pitt.edu/its-arch/group-ontology.html) identifies vocabularies in education to create a common language. Examples of reference sites/databases are THESAURI on ECHO (www2.echo.lu), with descriptions of 600+ thesauri and *A Web of online dictionaries* (www.bucknell.edu/~rbeard/diction.html), with links to 400+ searchable dictionary sites.

Each of these projects addresses some of the elements that a complete solution must address. **What sets SemWeb apart** is that it aims at addressing all of them in a comprehensive framework that maximizes knowledge integration, knowledge sharing, and knowledge reuse.

1. SemWeb presents the vision and basic architecture for a distributed intellectual infrastructure. It is a method that can be used in many contexts, including a potentially world-wide knowledge network that would sit on top of the mostly community-specific and geographically focused efforts and cross-link them. It is not conceived as managing one monolithic knowledge base.

2. SemWeb defines a comprehensive data schema or template for organizing ontological and lexical data. It provides for all conceivable types of information about concepts and terms and can thus be used by many different communities.

3. SemWeb provides direct access to resources, rather than just referring users to resources.

4. At the user’s option, SemWeb makes individual resources transparent by selecting the resources to be searched and integrating the results from querying several resources.

5. SemWeb can visualize concept relationships (from one or more ontological resources) by drawing a concept map, and it allows the user to navigate in such a concept map.

6. SemWeb provides an environment for the individual or collaborative development of resources, combining software support with optimal use of available resources.

7. SemWeb is interdisciplinary. It transcends any one community of users and contributors.

### 4 Related work and perspectives of the investigators

#### 4.1 Structure and construction of classifications and thesauri - Dagobert Soergel, College of Library and Information Services, University of Maryland

One important stream feeding into SemWeb is the development of classification schemes and thesauri for information retrieval. Dr. Soergel has been involved in this area both theoretically and practically for over 30 years. He has worked on the conceptual structure of classifications/ontologies and on data models for integrated thesaurus databases. He has developed TermMaster, a thesaurus development program that can store and correlate many thesauri; this program will serve as a starting point for tasks in this proposal. He has guided the development of the Alcohol and Other Drug Thesaurus, a major new thesaurus (National Institute on Alcohol
Abuse and Alcoholism, 10,000 descriptors in carefully structured hierarchies, 6,000 lead-in terms, 2,500 scope notes), with respect to both its conceptual structure and its presentation.

4.2 Development and integration of large ontologies and lexica - Eduard Hovy, Institute for Scientific Information, University of Southern California (ISI)

A second important stream feeding into SemWeb is the development of ontologies in the AI and NLP environment. Through his work in text summarization and text generation, Dr. Hovy has become involved in the development and integration of large concept taxonomies/ontologies. Over the past two years, he has collaborated with researchers at Stanford University, IBM, CYCorp, and ISI to create a large Reference Ontology from various existing ontologies, under the umbrella of the ANSI Ad Hoc Group on Ontology Standards (NCITS). This Reference Ontology is being constructed from ISI's 70,000-node ontology SENSUS (see http://mozart.isi.edu:8003/sensus/sensus_frame.html) and portions of EDR (Tokyo), CYC (Texas), MIKROKOSMOS (New Mexico), and other ontologies, using semi-automated alignment and validation techniques being developed at ISI by Dr. Hovy and his colleagues.

4.3 Information Visualization - Xia Lin, College of Information Science and Technology, Drexel University

Research indicates that information visualization helps to move the burden of processing large amounts of information from the user to the machine (Robertson, Card, & Mackinlay, 1993). Information visualization helps to condense, detect, and reveal "patterns" of the information to the user (Tufte, 1990). Information visualization also can utilize high-order statistics or semantic associations to convey content and meaning of text to the user (Wise, et al., 1995). Based on these principles, Lin (1997b) developed a web prototype called Visual SiteMap that, through the analysis of statistical patterns of words and links, visualizes all the links and the content structure of a web site and provides a new content-based access tool. In related research, Lin & Chan (1997) developed a visual interface, called Knowledge Class, that dynamically displays a collection of hierarchically structured terms on specific topics so that the user can select any level of detail of the hierarchy in order to interact with the information linking to those terms.

4.4 Conceptual structures for learning - David Jonassen, Department of Instructional Systems, College of Education, Pennsylvania State University

The use of conceptual structures is embedded in a constructivist approach to learning in K-12 and higher education pursued by Dr. Jonassen. The thrust of his research most relevant to this proposal is (1) the implementation of cognitive tools for learning (Mindtools) as knowledge representation formalisms that enable learners to represent their knowledge — what they have learned and what they are learning — in different ways and (2) the investigation of how the learner’s own cognitive structures interact with these Mindtools. Each of these Mindtools (e.g., databases, spreadsheets, semantic networking tools, expert systems, abductive reasoning tools, dynamic modeling tools, visualizations tools, microworl ds, hypermedia construction tools, and computer conferencing) engages learners in different cognitive and epistemic representations of what they know. Multiple forms of knowledge representation are necessary for complex and comprehensive mental model development. The more ways that learners can use to represent their knowledge, the more flexible and transferable that knowledge will be.
5 The proposed SemWeb architecture

5.1 The SemWeb template and the SemWeb interface

Underlying the entire system is a template for the arrangement of information about concepts and terms. Figure 2 gives an idea of the breadth of information to be accessed by a SemWeb system. Some of the slots pertain primarily to terms as linguistic entities; others pertain primarily to concepts. For many uses, a more simplified version would do, but with all the pieces of information required by any application (including special-purpose linguistic projects), the number of slots will be in the hundreds.

| Entry term, icon, concept (or a group of terms or concepts with common characteristics) |
| 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 (for educational and voice interface applications) |
| Word root and derivation from the root |
| Part of speech, inflection rules, and other syntactic information |
| 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/or how-to description (for functional concepts) |
| Usage notes, usage examples and quotations, familiarity and frequency. Explanation of subtle differences in meaning between related terms. Hyperlinks to texts in which the term occurs |
| Disambiguation rules. Rules on how to determine the proper meaning of a homonym |
| Detailed conceptual relationships (broader terms / hypernyms, narrower terms / hyponyms, parts / meronyms, whole / holonyms) and pointers to the concept's place in overall classificatory structures and conceptual maps. Display of the structural relationships among subordinate concepts. |
| Rules on combination with other concepts to form expressions. For verbs: case frame. |
| (Sub)language/(sub)culture/population group and audience level as a tag in every slot |

Fig. 2. A taste of the SemWeb template: Illustrative frame slots for organizing information about concepts and terms

The template focuses on information on individual concepts and terms. The user also needs views of overall conceptual/classificatory structures in various formats (linear listings, concept maps, etc.) for browsing and navigation, moving from a general overview to detailed sections. Some views are grand structures of knowledge, such as the great library classification schemes and more recent large-scale ontologies; others are local overviews and conceptual maps, such as
a table representing the relationships between the various specialized terms for horse (filly, mare, stallion, etc.) by age and sex, or a diagram of a causal pathway in biochemistry, or a representation of relationships among concepts in mathematics. SemWeb provides for collections of concept maps searchable by overall topic and educational objective served, as well as by the individual nodes. Each node (for example, a chemical substance or a process) could be expanded to show a definition and/or its place in another structure (such as a hierarchy of chemical substances). The user may want to see the same concept in several different structure views, each showing a different perspective.

The SemWeb template shapes the interface: The user starts with a blank 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. SemWeb 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. It also governs the system's internal workings by setting out the framework for organizing the system's knowledge about what information can be obtained from what source and how to search each source. The template also serves as an input form for sharing data. Finally, a frame hierarchy is one useful view of the internal structure of an integrated knowledge base.

The SemWeb template is to be developed based on existing standards and with the contributions from many communities. It will be maintained as a database that is open for additions and comments, providing a forum for discussion on the nature of data about concepts and terms. Purpose, origin, and status of each template element would be noted.

5.1.1 Special problem: Entity types and entity identifiers

The system deals with the following entity types: Character strings, terms and icons, concepts, and classes of any of these entity types that share common characteristics (such as a pattern of conjugation). The problem of assigning identifiers to the entity values of these various types is non-trivial, and a good solution is essential. The following examples illustrate 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.

EN-drill-n-OED-2 The English word represented by the character string drill, being a noun and listed in the Oxford English Dictionary as the second entry among the 5 nouns represented by this character string. The identifier EN-drill-n-W3-5 refers to the same word (as listed in Webster 3). The identifier FR-drill-n-HD-2 refers to a French word represented by the same character string.

EN-drill-n-OED-2.1 The concept corresponding to the first meaning of this word as listed in the OED (drill as tool). Drill as military exercise is EN-drill-n-OED-2.4.

These identifiers are unambiguous but not unique; a word or a concept has as many identifiers as it has sources. An integrated knowledge base or specific contributions from users (see below) would establish a correspondence between the different identifiers for the same word or concept.
5.2 The SemWeb resource inventory and metadata schema

Access to ontological and lexical knowledge bases in SemWeb is mediated through a master list in which each resource is described and analyzed following a comprehensive metadata schema. The user can keep a private resource inventory to be used in conjunction with the public one. The SemWeb prototype resource inventory will be open for collaborative maintenance, including attributed or non-attributed evaluations of specific resources (collaborative filtering).

The metadata schema is to be developed collaboratively with contributions from many communities. Sample slots are the domain, the kind of objects (terms, icons, concepts) covered, the kind of information given, function(s) supported, evaluations, address and access protocol, copyright and access fees and any other characteristics that are important for evaluating and selecting sources for any of the functions listed in Sec. 4.1. Specific characteristics for assessing the quality of a conceptual structure (including student-generated concept maps) include:

- Integratedness: Ratio of instances to concepts
- Embeddedness, interconnectedness: Proportion of nodes with multiple links
- Descriptiveness of links
- Parsimoniousness of links
- Ratio of the number of relations to the number of nodes.

Of particular interest are methods for deriving at least some of the resource characteristics automatically. This includes methods for the automatic or computer-assisted assessment of student-generated concept maps to be incorporated in the concept mapping tool developed as part of this proposal. Such methods are needed if concept maps are to be used in large classrooms for the assessment of student learning or for feedback to students.

5.3 The SemWeb knowledge exchange and integrated knowledge bases

To the extent possible, SemWeb will correlate data from different sources automatically. However, such automated integration is a very hard problem: homonyms need to be disambiguated; word senses may be delimited differently in different sources; specialty senses may not be covered in large standard sources; hierarchies may be constructed on different principles. The quality of data from some sources may be poor, and data from different sources are inconsistent. While SemWeb efficiently gathers and to some extent collates data from many sources, the user is largely on her own in interpreting the data. SemWeb will support further intellectual processing of the data gathered, resulting in a specially formatted Web page which can be shared with others, so that they can build on the intellectual effort already invested.

A SemWeb system may also include one or more integrated knowledge bases that incrementally accumulate and correlate data about terms and concepts from several sources, keeping track of the source for each piece of information. For many queries the user can restrict her search to an integrated knowledge base. Knowledge acquisition can be done centrally (with intelligent computer support). This integration can build on the Web pages submitted by individual users because these pages are structured in such a way that the information can be incorporated (with attribution) into an integrated knowledge base.

An integrated knowledge base can be extremely useful for collaborative development of new resources for specific purposes. SemWeb can compile a raw database of information about
concepts and terms in the domain. The editors can then work out the knowledge structures appropriate for their purpose and store the results in the integrated knowledge base properly tagged as belonging to the new resource developed. SemWeb can then extract a separate knowledge base to be used elsewhere or the user can define a special view on the integrated knowledge base that shows only items tagged as belonging to a specific resource.

5.4 The SemWeb software components

SemWeb is intended to be network-based (World Wide Web or Intranet). The SemWeb software will consist of a kernel for resource selection and data integration, plug-ins for accessing individual resources, and a tool for collaborative development of ontologies/classifications, concept maps, thesauri, and dictionaries; this tool will have an interface usable by students at all levels. It will be written in Java and C++, paying special attention to performance.

5.4.1 Integrated access component: Search and visualization (see Figure 1 on p. 2)

Kernel. Gets the user's request, selects the sources to be searched, integrates the information found, and displays it to the user. Various complications must be considered in the refinement of this component. For example, when the user wants all information about a concept, identified by a query term, SemWeb must also look for synonyms; but such synonyms are identified during the search, requiring a search in several phases. The unified format needs to support both ad-hoc search and massive data transfer for development of new schemes based on several sources. It should support a mode of operation that would enable an application, such as an NLP application, to access several resources as one virtual resource (that is, the application thinks it is dealing with one machine-readable dictionary, but the information comes in fact from several dictionaries that complement each other.)

Special plug-in modules for searching specific sources. The master list serves as a knowledge base for this software component. To the extent possible, searching specific sources will be accomplished through a general module that interprets encoded knowledge about sources. This will make it easier to make additional sources available through SemWeb. We will also consider defining specifications for resource-side plugins; this would enable a resource provider to provide access via SemWeb by installing on its server a plug-in that could understand SemWeb standard-format queries and send out standard format answers.

5.4.2 Knowledge base development and maintenance component

This component would serve users who need to develop their own resources or who want to create and maintain an integrated SemWeb knowledge base. It will include support for creating a pool of information from many sources, assistance with integrating information, support for input and editing of data (including support for entering and editing hierarchies in an outline format and for entering and editing concept maps, with an interface that can be used easily even by younger students), and support for collaboration among several users in the development of a lexical or ontological knowledge base, including students who collaborate on a concept map. It will build on Soergel’s work on data models for thesauri and his TermMaster program.


6 Research and development tasks to be accomplished in this proposal

The project will (1) develop the tools needed to build a SemWeb system; (2) build a prototype covering two or three domains to demonstrate the viability of the concept; and (3) study the use of the prototype for the many functions listed in Section 1, including learning. The goals of this project present a challenge. This section details how we intend to meet the challenge.

1 Develop a comprehensive SemWeb template for any kind of information on concepts and terms, bringing together existing standards for machine-readable dictionaries, ontologies, subject authority files, and classification data and results from a user needs analysis. (Basic template in Year 1, additions and changes Years 2 and 3)

Task 1.1 (Maryland). Analyze existing standards (examples in the references, a complete list to be developed for the proposal), resulting in a comprehensive list of all elements mentioned in at least one standard. Integrate the results of tasks 1.2-1.3.

Task 1.2 (ISI). Analyze the literature on the structure of machine-readable dictionaries and ontologies to discover any additional data elements to be included. Analyze NLP projects to see what kind of information they used.

Task 1.3 (Maryland). Same for thesaurus, authority, and classification data.

Task 1.4 (Penn State). Same for data represented in concept maps.

2 Develop a comprehensive metadata schema for resource description and evaluation as the basis for linking resource inventories into a SemWeb system and for exchanging descriptions and evaluations of resources.

Task 2.1 (Maryland). Compile a list of data elements to be included in the metadata schema from available schemas, to be further expanded by Task 2.2. (Year 1)

Task 2.2 (ISI, Maryland, Penn State). Work on the problems of describing and evaluating (measuring) resources (dictionaries, thesauri, classifications, ontologies, concept maps). Bring together ideas on describing/evaluating ontologies in AI, thesauri and classifications in IR, and concept maps in education, building on work done by Hovy, Soergel, and Jonassen, and good literature. The result would be a list of variables or a template, with instructions on how to fill in the slots. This includes developing methods for computer-assisted scoring of student-generated concept maps. (Years 1-3)

3 Develop the integrated access software component: Search and visualization and specifications for plug-ins that link to specific resources. (Basic version Year 1, refinements Years 2 and 3)

Task 3.1 (Drexel, Maryland). Analyze the literature on implementing lexical resources on the Web and on user interfaces to thesauri and on the visualization of relationships in thesauri, classifications and ontologies. Analyze actual systems.

Task 3.2 (Drexel, ISI, Maryland, Penn State). Write interface specifications.
Task 3.3 (Drexel, Maryland). Develop a method for mapping from any lexical resource data structure into the SemWeb template. Apply the method to sample systems.

Task 3.4 (Drexel). Write software for searching lexical resources on the Web and bring the results into a unified format on the SemWeb server. Write the kernel and some plug-in modules for resources that have priority for inclusion in SemWeb.

Task 3.5 (Drexel). Write interface software.

4 Develop software for computer-supported collaborative ontological engineering.

Task 4.1 (Maryland, ISI). Analyze the state-of-the art of methods for developing thesauri, classifications, and ontologies, including existing software. Review applicable computer-supported collaborative work literature. (Year 2)

Task 4.2 (Penn State, Drexel). Look at a number of concept mapping tools, get feature list, develop tools students can use alone or collaboratively, with support through SemWeb (for example, by providing definitions to students). Provide special interface features for teachers and students. Teachers should be able to edit representations and add new ones, supported by the SemWeb system in general. (Years 1 and 2)

Task 4.3 (Maryland, ISI). Develop methods for automated comparison and integration of different dictionaries, thesauri, classifications, ontologies, and concept maps. Particularly difficult problems here: (1) comparison of hierarchical structures and (2) comparison of NLP dictionaries constructed with different linguistic perspectives. (Year 3)

Task 4.4 (Maryland, ISI, Drexel, Penn State). Develop specifications for ontology development software, considering use of existing sources, automatic methods, and collaboration. (Year 2)

Task 4.5 (Drexel). Develop and test software. (Year 2)

5 Analyze several domains — users, uses, and resources — and work with domain organizations to prepare for the prototype and develop a resource inventory. Possibilities include mathematics (where SemWeb could eventually be extended from including just definitions of mathematical concepts to include mathematical ideas, possibly establishing a venue in which advanced ideas could be presented in a way that is accessible to mathematics students); physics, chemistry, and biology (where concept maps can be very helpful); alcohol and other drugs (AOD, a highly interdisciplinary field ranging from neurochemistry and brain imaging to psychology, sociology and law); research methods, tests and measurement across disciplines. (Year 2)

Task 5.1 (ISI). Analyze users, uses, and resources in several domains from a knowledge-based systems and natural language processing perspective.

Task 5.2 (Maryland). Analyze users, uses, and resources in several domains from an information retrieval perspective.

Task 5.3 (Penn State). Analyze users, uses, and resources in several domains from an education perspective.
6 Implement a prototype WWW SemWeb server and publicize it.

Task 6.1 (Maryland, Drexel). Set up the SemWeb server with a number of general resources, such as Dewey Decimal Classification, Library of Congress Classification, Library of Congress Subject Headings, Yahoo classification, WordNet, CYC upper ontology, SENSUS (ISI ontology), Alcohol and Other Drug Thesaurus (for its hierarchies and scope notes in many domains), UMLS (as an example of an integrated knowledge base, even though domain-specific), Webster’s, Oxford English Dictionary (possibility of linking in CD-ROM or use as example of how to deal with user fees) (Years 1 and 2)

Task 6.2 (Maryland, ISI). Develop a broad resource inventory to be put on the server. (Year 2)

Task 6.3 (Maryland, Drexel, ISI, Penn State). Based on the results of Task 5, select two or three specific domains (in different fields, but also related, with good possibilities for applications) and link in resources from that domain so that they are accessible through the server (this involves programming the plug-in modules). (Year 2)

Task 6.4 (Penn State). Compile a collection of structural representations of knowledge in a selected domain. (Years 1 and 2)

7 Pilot testing and use analysis (Year 3)

Task 7.1 (Maryland, perhaps a PhD thesis). Analyze the use of the SemWeb server (who uses it for what purpose with what effect, questions asked, information contributed).

Task 7.2 (ISI). Carry out an NLP project with SemWeb support and analyze experience.

Task 7.3 (Penn State). Use the software for SemWeb-supported concept mapping in the classroom and test effects on students’ learning. Apply methods for computer-assisted scoring of student-drawn concept maps.

Task 7.4 (Penn State). Do a pilot project for studying the use of concept maps for instructional materials development and for communicating complex interrelationships to students. This would be supported by the collection of concept maps on the server.

Task 7.5 (all, emphasis Drexel). Usability studies (accompanying software development).

7 Projects that could build on SemWeb. Examples

This section gives a small sample of project ideas to further illustrate applications of SemWeb.

Value-added searching with SemWeb on the Web, in digital libraries, or in bibliographic databases. This integration could be done on a server (such as AltaVista or other large Web search system or OCLC’s First Search) or on a client through a browser plug-in. By linking directly to SemWeb, a search system could make a vast array of ontological and lexical resources conveniently available for improving searching and document comprehension. In query formulation, the user could click on a search term (word or phrase) to be presented with entries for its various meanings, with a definition and listing of synonyms, broader, narrower, and related terms for each meaning. Or the user could browse hierarchy displays or other
visualizations of concept relationships drawn from suitable sources. By simply clicking on a term in that display, the user could copy that term into the search window. Alternatively, the user could opt for automatic query expansion for which the system would use synonym and hierarchical relationships collected from one or more sources. In understanding texts found, the user could click on any term in the text and be presented with a definition. The system would use the most obvious source first (this requires some information about the user, such as reading level) and consult other sources only if the term cannot be found in the first source or if the user wants more definitions. In a more sophisticated version, the system could figure out which of the meanings of a term is most likely in the context and present that definition first, considering also the user’s background.

As a special case of such integration, SemWeb could support the Dublin Core of basic metadata for document-like objects on the Web. The Dublin Core provides for indicating the scheme from which a subject term is taken through an acronym. The SemWeb server could be enhanced to include in its inventory all schemes that occur in Dublin Core records so that the user would be a click away from information about a scheme and access to it.

**Semantic analysis of classifications for better user access.** Creating a faceted classification of elemental concepts to analyze the classes in Dewey Decimal Classification, Library of Congress Classification, and the Library of Congress Subject Headings to implement a conceptual index (Soergel 1972, 1990). This would both use SemWeb and enhance the usefulness of the server.

**Automated understanding and comparison of definitions.** Frame hierarchy of formal definitions. The user would be greatly added by a system that could understand definitions and give a comparative analysis of several definitions. Deriving formal definitions and arranging them in a frame hierarchy (possibly with human editing) (Slator 1990) would support this purpose and could be a powerful tool for modeling complex dictionary entries and for capturing subtle differences in meaning, especially across languages. This in turn might be used to improve the quality of natural language processing, especially text generation and automated translation.

**Links to tools that assist in deriving terms and concepts and relationships (including translations) from corpora.** A variety of such tools exist, and knowledge about them and quick access to them would be quite helpful in the construction of dictionaries and ontologies.

**Links to NLP tools.** SemWeb could be complemented by links to various inventories of NLP (natural language processing) tools such as part-of-speech taggers and syntactic analyzers or, even better, incorporate these inventories. Then there could be links between tools and the specific ontological and lexical resources they work with. The SemWeb environment could be enhanced to provide easier access to the tools themselves, such as simple click to download or even run on a server machine with one’s own data. SemWeb could also provide an environment for sharing experiences in using such tools.

**Study of the interaction between users’ cognitive structures and conceptual structures presented by a tool such as SemWeb**
Dissemination plan

Dissemination is essential to this project for three reasons:

1. We want several communities to make comments and suggestions on the design of SemWeb.

2. We want the SemWeb server widely used.

3. We want to encourage others to undertake research and development that would make use of the possibilities offered by the SemWeb environment and extend that environment.

We will use all the usual means of dissemination, keeping in mind that we need to communicate to four different audiences: Linguistics / Natural Language Processing; AI /Knowledge-Based Systems / Data Modeling; Information Studies / Information Retrieval / Librarianship; Education

• Presentations at conferences. Workshops and round tables at conferences (to solicit and collect suggestions and comments). In Year 2 or 3 we might consider proposing a Symposium in the AAAI series on a related topic, such as Intelligent Integration of Ontologies.

• Web-based dissemination. This will include a Web site (on the same machine as the SemWeb server) with general information on the project (different pages tailored to different audiences), progress reports and other project documents, and links to related sites. We will make a special effort to have this site, as well as the SemWeb server itself, listed on many reference sites directed at each of the audiences, as well as on many library reference sites. The site will invite comments and suggestions.

• Notices in journals, newsletters, news groups, and mailing lists, both print and electronic.

• Journal articles that will document the results of the project.

• Distribution of technical reports, both electronically and in print.

• Personal correspondence to and other contacts with individuals and institutions who might be encouraged to start their own related projects.
Institutional commitments as to space and equipment

The four participating institutions (University of Maryland, University of Southern California Institute of Scientific Information, Drexel University, and Pennsylvania State University) commit to providing the space and the communication and computer equipment necessary for the project staff to carry out their tasks. (NSF funds are requested in the Drexel budget for the SemWeb server, which will be used primarily for making a SemWeb system available for public use but also for program development and for project communication.)
Performance goals

Year 1

Basic SemWeb template (Task 1)
Basic metadata schema for the inventory of ontological and lexical resources (Task 2.1)
Basic version of the SemWeb kernel, the search and visualization software for a SemWeb WWW server (Task 3)
Plug-ins for some general resources (Task 3)

Year 2

SemWeb server running with access to general resources (early in Year 2) (Task 6)
Report on the analysis of needs and resources in several domains with selection of domains to be chosen for the prototype (June 30 of Year 2) (Task 5)
Database of concept maps available through server (June of Year 2) (Task 6.4)
Broad resource inventory (Task 6.2)
Plug-in modules for resources from the selected domain implemented on the server. (Task 6.3)
First draft of a report on methods for describing and evaluating of ontological and lexical resources (Task 2.2)
Software for managing SemWeb’s own integrated database and for collaborative development, including concept mapping software (Task 4)

Year 3

Report on automated comparison and integration of different schemes and corresponding additions to the development software (Task 4.3)
Software refinements
Additional resources accessible through the server
Report on the use of the server: Amount of use, types of users, how do they incorporate the SemWeb answers into their work (Task 7.1)
Report on the usefulness of SemWeb in an NLP project (Task 7.2)
Report on students’ learning from drawing concept maps, alone or in groups, and on student assessment through concept maps (Task 7.3)
Report on the use of concept maps in instructional material development and in presenting complex subjects (Task 7.4)