Dagobert Soergel ds52@umail.umd.edu www.clis.umd.edu/faculty/soergel/ July 1998

A digital library model with rich semantic structure

Summary

This project will develop a blueprint for an enriched digital library — an information space with rich semantics and functionality, including well-structured concept hierarchies — that integrates information-seeking with user tasks and supports the management of private information spaces and cooperative work, especially user enhancements to the database and collaborative authoring; it will build on whatever already exists to implement a part of this blueprint using the 20/80 rule (20% of the effort achieves 80% of the functionality). The enriched digital library model will specify an information structure and an integrated set of functions with links to where work on each function has been or is being done. Building on existing work, and expanding on it, the project will develop a digital library toolbox with basic (as opposed to fancy) implementations of the most important of the functions specified in the model. This toolbox will be useful to anybody who wants to build an enriched digital library in a content area.

This project is part of a larger research agenda which aims at applying the enriched digital library model in a field laboratory in the socially important area of Alcohol and Other Drugs (AOD) / Substance Abuse and Prevention, assembling and interlinking information from several government and private agencies, and at using this field laboratory to study how information system content and functionality and cognitive processes of users work together in completing tasks, to find out what content and what function best support users.

To support individual and collaborative work, the digital library model includes a public information space and private information spaces for individuals and work groups. Users have many tools to manage their private space, can search their private space separately or integrated with the public space, and are encouraged to contribute documents, annotations, and links.

The information structure envisioned covers several object types (such as concepts, documents, projects, persons, organizations, and queries) and link types (such as deals with, defines, produces, includes, cites). It emphasizes concept hierarchies to guide exploration.

The functions or tools envisioned are divided into search tools, especially a tool for browsing concept hierarchies, a search form for searching the system and external databases, thesaurus support for searching, and a glossary tool; document creation and editing tools, supporting collaborative authoring and annotation and link management; and document processing and link generation tools, including a document segmentation tool to structure plain text, a citation index tool, a summarization tool, and a document categorization and indexing a tool.

The approach of the project will be to create an inventory of existing tools (often protoypes), select the best of them, and adapt and repackage them into portable production versions and to create entirely new tools only where nothing already exists. All tools will be tested for functionality and usability. The project will assemble a small prototype digital library as a test environment and will also do some initial studies of the use of such a system.

A digital library model with rich semantic structure

1 Introduction

This project will develop a blueprint for an enriched digital library — an information space with rich semantics and functionality, including well-structured concept hierarchies — that integrates information-seeking with user tasks and supports the management of private information spaces and collaborative work, especially user enhancements to the database and collaborative authoring; it will build on whatever already exists to implement a part of this blueprint using the 20/80 rule (20% of the effort achieves 80% of the functionality). This will form the basis for a companion project, the construction of an Alcohol and Other Drug (AOD) Digital Library as a field laboratory for studying the use of such a facility.

We start out with two vignettes that illustrate the ultimate functionality we aim at, realizing full well that not all of this can be accomplished in this project.

Scenario 1. The director of a drug-free community coalition is faced with developing a prevention project and the funding for it. Signing on to the AOD Digital Library, she begins by browsing the prevention section of the thesaurus hierarchy to get a structured overview of various prevention approaches. From the thesaurus scope notes, some of these approaches seem particularly applicable to her community, so she follows the links to more in-depth explanations. She returns to the thesaurus and follows a link from *prevention through education* to a funding program announcement. She opens the guidelines for submitting proposals to this program and copies a proposal template into her private space (shown in another window) and fills in some text and copies some text (which is transferred with the proper source). From the program announcement, she follows a link to projects funded previously and further to project reports and evaluations. She comes across the unfamiliar term *triangulation* and clicks on it to see the thesaurus entry, which gives an explanation and the hierarchical context. In another document she highlights the phrase *prevention program evaluation* to initiate a search in the system and one external database. She copies three references with abstracts to her private space. (Later she will return to these, select one for detailed reading, and add more notes and quotes to her emerging proposal.) Returning to the program announcement, she follows a link to relevant research, selects some articles to read, and adds more material to her outline. One of the papers compares the effectiveness of several prevention curricula. She follows a link to the curriculum that came out on top and from there finds further reviews. She also finds some discussion of resources required. She needs some more data — namely, demographics of her community and funding sources for the required local match — so she initiates searches in two external databases, incorporating the results into her proposal. Now she completes the first draft, including the text itself and annotations that explain why a piece is included or why certain language is used. Before submitting the proposal, she emails two board members and a city staff member for comments, giving them access to her private space. The three people read the draft and add their annotations, including suggested wording. The director now revises the

draft, requests the final document in PDF format, links to the agency's submission system, and sends off her proposal.

Scenario 2. Three researchers are collaborating on a paper on the genetic basis for alcoholism. They use all the functionality discussed above and more. They have developed a list of issues and need to collect material on each issue. As they start assigning texts to issues, the system's automatic classifier trains incrementally so that it can assist in the assignment of new text sections to issues.

One of the issues is the incidence of alcoholism in families, and one of the researchers starts an exhaustive search on the subject. He finds *familial alcoholism* in the thesaurus but not many cross-references. He looks to see whether somebody else did such a search and posted it publicly. He finds a search form, which includes the terms *twin study* and *adoption study*, among others. The old search is not quite what he needs, so he modifies the search form and runs a new search. The system acquires the documents found and prepares summaries; it also clusters the results, storing it all in the group's space. The researcher examines some of the clusters himself and, via email, asks his colleague to examine the others, because they need different expertise. The colleague sees an interesting paper available in PDF format. She tells the system to acquire it. The system downloads the document (preserving the original address), divides it into meaningful segments, adds XML markup, analyses the bibliography, adds links from each in-text citation to the corresponding entry in the bibliography, formats each bibliography entry into a bibliographic record to be added to the system's bibliographic database, searches the Web for the full text of each database and adds any URLs found to the bibliographic records, and finally assigns the document as a whole and its individual sections to the appropriate issue(s). The researchers can now examine all the documents or document sections under each issue. (If they revise an assignment, the classifier learns and will do better next time.)

One of the researchers, while reading an important paper, follows one of the automatically created citation links forward in time, is taken right into the proper context in the citing document, and notes a contradiction between two documents; he creates a public *contradiction* link so that another user, reading only one of the documents, becomes aware of the contradiction. Reading further, he has an idea that builds on the author's argument and writes it down in a public annotation. For another document, the researcher makes an annotation giving his personal evaluation, which he keeps private in the work group space. (He could also keep it entirely to himself in his individual space). These researchers require lots of data collected in other studies. They create a database and enter much of the data by simply converting tables given in the relevant publications, supplying column and row headings from their classification of variables. Other data they obtain by sending queries to external statistical databases (such as FedStats ─ http://www.fedstats.gov) and to genome databases. The system takes care of mapping the query to the proper format and mapping the results returned so they can be easily added to the database. When they publish their paper, the researchers might make their database (which has a lot of added value) publicly available.

The researchers are now ready to begin writing their paper, perhaps initially working on separate sections. The whole paper is stored in their group space, so each of them can see it.

Later they will annotate each other's pieces or work on the text together during a conference call.

2 Objectives

The immediate objective of this project is to build a tool box for us and others to use to create digital libraries with rich semantic structure and functionality. The longer-range objective of the research plan which motivates this project is to build an Alcohol and Other Drug (AOD) digital library and use it as a field laboratory for the study of the use of such a facility and the cognitive processes of users as they interact with it.

The objective of this proposal can be further broken down as follows:

- to develop a blueprint for an enriched digital library;
- to implement and test a basic version of this blueprint, incorporating existing tools wherever possible, with emphasis on simplicity and basic functionality in a toolbox that can be deployed easily;
- to make available the functionality of the system with a small sample collection, including the Alcohol and Other Drug Thesaurus;
- to test the usability of the system and get a glimpse at the ways in which users interact with the system and use it for their work.

To achieve its objective, this project needs to build a total environment. This is feasible only by maintaining a tight focus on the main goal — making rich semantic structure available to the user. Therefore, we will use the simplest means that will accomplish this objective and forego many features that should be added later: advanced visualization; advanced multimedia capabilities (we simply use what is available on the Web); digitization (we use materials already available in machine-readable form); cross-language retrieval and translation; copyright and payment administration (we will deal with government and other publicly available documents).

A companion proposal, to be developed jointly with the Center for Substance Abuse Research (CESAR), University of Maryland College Park, will deal with using this toolbox to set up a digital library in the field of Alcohol and Other Drug (AOD) / substance abuse and prevention (AODLib). AODLib would be built around the AOD Thesaurus, which provides a conceptual map to this interdisciplinary field. It would combine and interlink documents from several Federal agencies — National Institute on Alcohol Abuse and Alcoholism (NIAAA) and National Institute on Drug Abuse (NIDA), both in NIH, and Center for Substance Abuse and Prevention (CSAP) and Center for Substance Abuse Treatment, both in SAMHSA — and from the substantial collection available on the CESAR Web site; while many of these documents are now available on well-designed agency Web sites, their integration into one structure and the added functionality would significantly improve access and usability. AODLib would also support authoring, especially the creation of agency documents, which often requires contributions, comments, and approvals from multiple parties. It would emphasize links from research results to prevention goals to enhance information transfer from research to practice. This digital library — and others applying the toolbox — would provide a field laboratory for studying the use of information in carrying out important tasks and the contribution of the various content and functional elements.

3 Approach

We will develop a definition of the structure and functionality of an enriched digital library model on the basis of an analysis and systematization of selected literature on digital libraries, hypermedia, and computer-supported cooperative work. We will do an extensive survey of what tools exist now, their status and availability; the result will be a systematic catalog with brief descriptions and references to demo sites and literature. We will then match our functional specification against this catalog to determine which functions can be accomplished with existing components (with adaptions as needed) and which functions require new components. From this match we will estimate the effort for each function and develop a work plan. We will take the best of digital library technology, develop it further, and bring it to practical application. This process — already started in this proposal — will result in a practical, portable, publicly available Web-based system for a digital library environment rich in semantic structure and functionality.

We expect to advance the state of the art in many of the tools we adapt or create, particularly in thesaurus interfaces, document segmentation, computer-assisted indexing, and the use of typed links in a large-scale system.

4 The vision

After a brief discussion of users and uses (Section 4.1), this chapter describes a preliminary system design which further specifies the work to be done. Central to our system design is support for the users' work and for information sharing and collaboration. This requires that users be able to contribute to the public information space and also have private information spaces in which they can store queries, documents, annotations, links, subject indexing information, or whatever; Section 4.2 deals briefly with this recurring theme. Section 4.3 explains the information structure that is the backbone of the system. Section 4.4 addresses the plug-in system architecture; Section 4.5 then outlines thesaurus interface and search functions. A number of tools are needed to create and maintain an enriched digital library, both public and private information spaces; these are listed in Sections 4.6 and 4.7, divided into interactive editing tools (Section 4.6) and more batch-oriented document processing tools (Section 4.7). The lines between Sections 4.5 - 4.7 are not sharp; there is considerable overlap.

For simplicity of language we describe the system as if it existed, even though it is just in the idea stage. For references see the bibliography, which is arranged by the proposal outline.

4.1 Users and uses

The tool box to be developed is designed to support users at all levels. By way of example, we list here the intended users and uses of the proposed Alcohol and Other Drug Digital Library.

- Researchers in all areas of the AOD field for ease of access to scientific literature and to data, especially reports from prevention and treatment projects (data to study the effectiveness of prevention and treatment measures).

- Practitioners on the front lines of treatment and prevention for planning, implementing, and evaluating prevention and treatment programs (including writing proposals) — information on research results bearing on their projects, reports of similar projects, and information on funding opportunities.
- Policy makers and planners of program announcements for research and prevention and treatment programs — information on problem situations, approaches that worked, results of health services research..
- Educators at all levels and authors of prevention and treatment manuals.
- Students and the general public — validated information on Alcohol and Other Drugs presented in an attractive and easily understood format.
- For all users: Sharing annotations and links and other enhancements to the database, and efficient preparation of documents and collaborative authoring — from the preparation of government-issued documents that require participation by many people (scientists, editors, project manager, sometimes advisory boards) to the writing of a student paper.

4.2 Public and private information spaces

The envisioned system is divided into a public space and private spaces for individuals or work groups. Any individual user or work group can interact with the system as a unified whole of the public space and the applicable private spaces.

The **public space** includes objects and links acquired using the tools described in Section 4.7 and shared objects, annotations, and links contributed by users. Some of the content of the public space is validated by people with proper authorization.

Private spaces are provided to individuals and work groups for storing queries, query results, bookmarks, annotations, links, notes, documents acquired from elsewhere, documents being worked on, etc. All the tools described in Sections 4.5 - 4.7 are available to users for acquiring and organizing information in their private information spaces. At the user's option, the system can use the private space to remember user interactions, such as documents looked at, and build a user model for improved interaction with that user.

The system keeps track of the author of each piece of information; objects and links are clearly identified as to author/origin. Links are shown in different colors depending on their status: "Official" links, shared links, and private links. Users can also register their support for a given piece of information (a document, annotation, link, etc.); this feature can be used for collaborative filtering. If a user does not log in, authorship or support is shown as *anonymous*.

A system open to contributions and enhancement by users needs some amount of regulation and security administration, which needs to be worked out.

4.3 **Information structure** (Figure 1)

The proposed digital library system goes well beyond dealing with text and images. It includes many object types and a rich network of typed links — from concepts and people to organizations, projects, and documents; from queries to any object retrieved and any object selected by the user; from projects to documents; from document section to document section; from research results to prevention strategies; from a prevention RFP to the most applicable research results. This structure supports the user in navigation, retrieval, and information use.

The system stores information in the form of structured documents and as databases (thesaurus database, project database, bibliographic database), with embedded or separately stored links. Documents describing objects such as persons, organizations, or projects are marked with the type of object they describe.

Documents in the system have a fine-grained structure. They are linked through hypertext links that originate in a specific place in a document and end in a specific place in the target document. Where applicable, documents are structured using one of several standard schemas (for example, purpose, methods, results, discussion) with a table of contents. Citations in a document text are linked to the corresponding entries in the document's bibliography and to the bibliographic database. The reverse links go from a document to later documents that cite it (citation index link); the link destination is a specific place in the citing document.

A key component is a thesaurus with well-structured concept hierarchies (such as the Alcohol and Other Drug Thesaurus) to help the user get oriented in an unfamiliar domain (particularly important in interdisciplinary areas). A thesaurus term can be linked to a document section that gives background information beyond the definition given in the thesaurus. There is also a special document type, *concept map* (for example, a diagram showing structural relationships among concepts or a diagram showing the causal influences in a set of variables); a thesaurus term can be linked to the specific location of the term in a concept map.

The system includes several internal databases:

- A bibliographic database with records for the documents in the system, documents mentioned in footnotes or bibliographies, and documents from other sources. The bibliographic records include the usual metadata (Dublin core at the minimum), plus document type designations and intended audience / level (both important elements for retrieval or arrangement of retrieval results), and links to the full text where available.
- A database of research projects and of "implementation" projects (such as prevention projects or education projects); searchable fields, include thesaurus descriptors. A project is linked to (and thus searchable by) the RFP or program announcement under which it is funded. A project is linked to its home page and to documents originating from the project. This allows a user who sees an interesting document to navigate to the project it came from and from there to further documents originating from the project.
- Databases for persons and organizations, with links to their home pages where available.

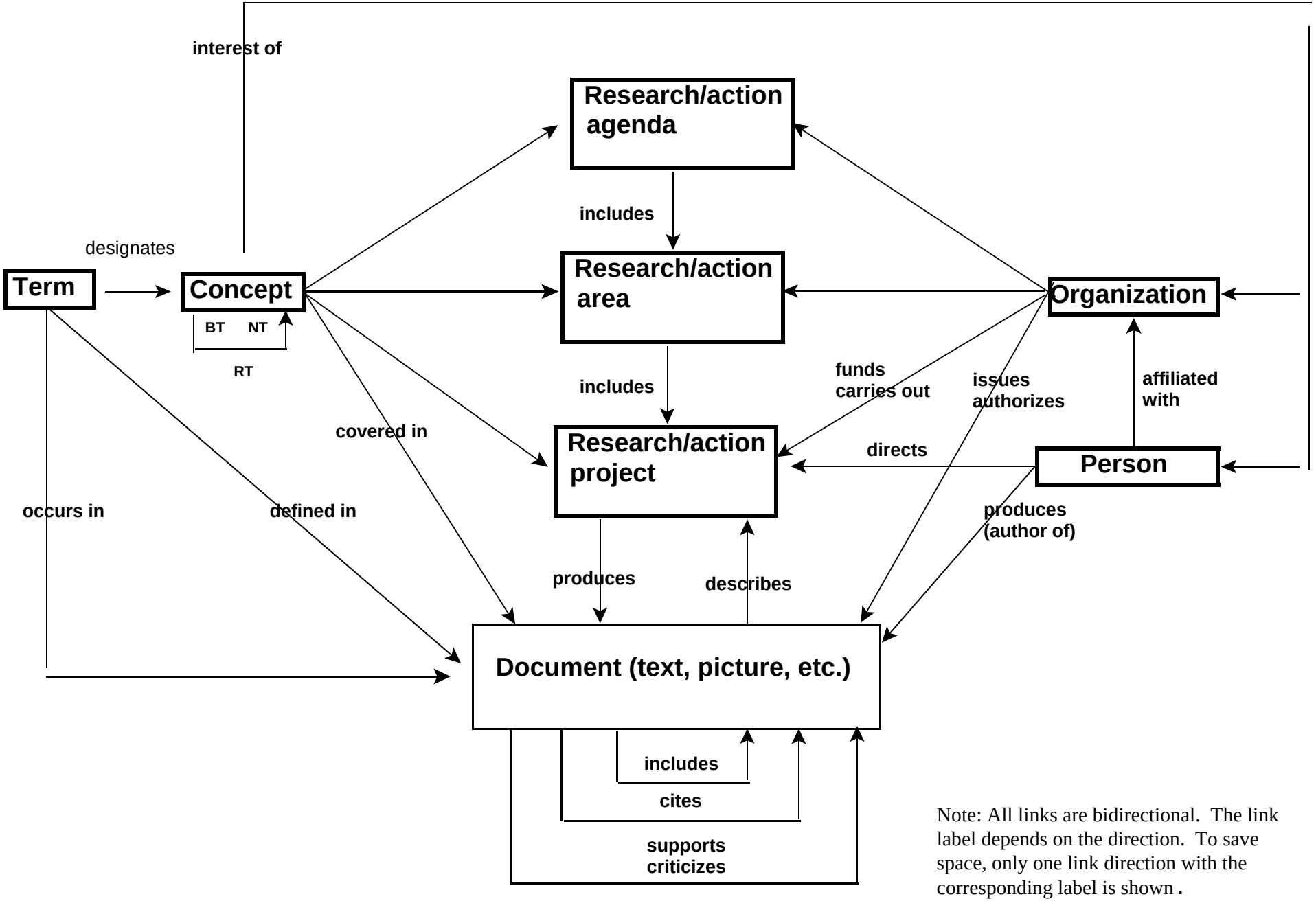

Figure 1. Information structure of the proposed digital library system

Queries can be stored as objects and can be indexed, annotated, and linked to items retrieved and (with a different link type) to items selected.

From the user's perspective, links are typed. The actual links can be untyped, as is most common on the Web now, or typed, as possible in XML. In either case, link types are specified in a systematic fashion in the anchor text, which indicates the nature of the target and the nature of the relationship, both drawn from fixed (but expandable) lists, with the possibility for further specification of the relationship by additional text. All links are bidirectional (reverse links generated automatically). Links can be annotated. A clickable symbol marks an annotated link.

The management of links needs to support private links and user-entered links to and from documents external to the system, as well as recording the authorship of links. This requires that at least some links be stored in a database, rather than in the documents themselves.

4.4 System architecture

The system uses a plug-in architecture informed by the Stanford information bus. Individual tools or components either follow a specification for interfacing with the system or an interface module that converts incoming and outgoing data to the proper format. One key element will be a flexible format for the representation of documents with which many components can work.

4.5 Search and other user-system interaction functions

Thesaurus interface. A simple but functional interface for browsing the concept hierarchies and for searching for terms whose synonym set contains the search word(s). (See the appendix.)

Glossary function. An expanded glossary function lets the user select a term (word or phrase) and display the thesaurus entry for the term or search the thesaurus for terms containing the word or phrase; initiate a free-text search for the term; put the term on the clipboard; or insert a link or annotation. This function could be expanded to search external vocabulary servers.

Search functions. The full text of the entire library is searchable, with the option of fielded search and search exploiting the document structure. The search can be extended to include databases that are tightly coupled with the digital library (especially the database of projects), as well as external databases.

The system displays a **search form** for the specification of field values and search logic, search term expansion using the thesaurus, search domain, and instructions for the sorting of results. The search form becomes part of the private space; it can be saved and indexed, and it can be shared by posting it to the public space. (Sharing forms for successful searches, preferably with the links to objects judged relevant, is a good way of helping other users.) The search form can be invoked from multiple places: clicking on a term in a text; clicking on a page representing a document, person, or project (brings up a search form for a document, person, or project, respectively, and fills in the form to search for similar objects of the type); browsing through (or searching for) previously used search forms, public or private.

The search domain can consist of any combination of internal and external databases / search services. External databases can be of various types, such as bibliographic databases; vocabulary databases; databases of people, organizations, or projects; and numeric/statistical databases. The system translates the search form into the search syntax of an external database; at the user's option, it will connect to the external database and show its native search form filled in so that the user can modify it. The system reformats the search results received from external databases and integrates the results from multiple like databases (particularly challenging if each of multiple databases contributes a different type of data). This functionality needs several tools: A sophisticated **query mapping tool** (most query translation tools map to the simple search in the target system), a **result mapping tool** (based on Z39.50), and a **duplicate detection tool**.

The user can invoke the summarization tool (see Section 4.7) to work on the search results. The system obtains the full text of documents retrieved and summarizes them.

The user can specify the arrangement of search results, sorted by any field (including URL) or arranged in clusters (arranged sequentially or in a two-dimensional map), using an **arrangement tool** that can be applied to any set of objects.

Hierarchical views for browsing. The system supports hierarchical views of an information space from various perspectives (provided the data to generate the view are available). For example, documents and other objects could be indexed using a hierarchical classification (which would include precombined descriptors), to produce a "directory" view of an information space. (Indexing might be accomplished by dragging an object into a "folder".)

Enriched document display tool. Inserts links and/or annotations stored in a database into the document text and passes this enriched text to the browser for display. The system also allows viewing a composite document (consisting of several small documents linked by inclusion relationships) as a single unit.

Graphical browser for browsing any kind of structure, such as the thesaurus structure as a semantic network, the internal structure of a single document, or the relationships within a set of documents.

4.6 Reading tools and document creation and editing tools

Reading, annotating, and creating new documents are highly intertwined activities and are therefore covered in one section. The system makes creating new documents easy in order to encourage users to use the system and thus increase the database. (For example, the system works with the user's preferred word processor). Documents created on the system can be better structured and linked than preexisting documents processed into the system.

Reading, annotation, and linking tool. While reading, the user can search the text for any word or phrase or for a word or phrase that designates a concept in a given category (for example, geographical places, anatomical sites, chemical substances, or genes) or request that such terms be shown in a given color. The user can highlight portions of the text for copy; the source will be attached automatically. The user can add annotations and links to internal and external

documents; these annotations and links may be shared or private (see Section 4.2). The system supports uniform anchor text to specify link types by providing menus of link types and object types. Ideally, the system would support anchoring links and annotations at any location within a document and targeting links to any location in the targeted document. In practice, there will be limitations, more stringent for external documents than for internal documents.

Authoring tool. The system supports collaborative authoring of structured documents (SGML or XML coded) with version control and authenticated approval. It provides a number of document templates to choose from. The text of an existing document can be included in the new document being worked on through an *include here* link.

Print tool. The system supports creation of print documents (and PDF or PostScript files) from SGML or XML coded documents or several subdocuments linked together.

Specialized tools used in creating documents

- **Bibliography tool**. Creates a bibliography from links to the bibliographic database or to bibliographic records found in a search and formats the bibliography to any of several styles.

- **Glossary tool**. Creates a glossary based on definitions found in the thesaurus. Terms to be included in the glossary can be marked manually or selected automatically (assumes that the thesaurus includes information about which terms are unfamiliar to what audience). If a glossary term is not in the thesaurus or does not have a scope note, the user fills the void for the present glossary and the system sends a suggestion to the thesaurus editor.

- **Index and concordance tool**. Prepares an index based on marked entries or based on automatic selection. At the user's option, standardizes terminology using the thesaurus and includes references from non-preferred terms.

Status management tool. The system supports "promoting" documents, annotations, links, etc., from private to shared. A user can also retract authorship; what happens to the object or link after that is a complex question, since other people may have supported or validated it.

4.7 Document processing and link generation tools

Document acquisition tool. Acquires documents in any format and converts them to the system format for easy processing by other tools. This includes scanning, but we will deal only with documents that are available in machine-readable form, particularly with documents that can be acquired from the Web (including PDF and PostScript). The resulting version of the document may then be further modified as required. A link to the original document will be maintained.

Document segmentation tool. Takes large unstructured documents and divides them into sections (chapters or individual papers, subchapter, etc.) based on (1) headings or other format

characteristics and (2) content analysis, applying a standard schema where possible The tool will work fully automated or with editor intervention as needed.

Citation link and citation index tool. Establishes links from citations in document text (1) to the corresponding entries in the bibliography and (2) to the corresponding record in a bibliographic database..

Bibliographic citation parser. Parses bibliographic citations found in documents into a bibliographic record to be included in the bibliographic database.

Metadata generation tool. Extracts metadata from a full document to create a basic bibliographic record.

Summarization tool. Creates a summary of a document. There is much work, going back over 30 years, on this task.

Automatic/computer-assisted document categorization and indexing tool. Assigns index terms automatically, either terms drawn from the document or descriptors taken from a thesaurus, or terms from the private classification of a user expressing her specific interests. Document categorization will determine, among other things, the type of document, its audience, and its reading level. This tool can be trained with a training set of documents carefully indexed by competent indexers. This tool also has an interface to assist a human indexer, possibly using the index terms derived automatically as suggestions. There is much work on both of these aspects.

Document finder. An agent that starts from a bibliographic record, finds the full text of the document on the Web, and puts the URL into the bibliographic record.

Home-page finder. An agent that finds personal, organizational, and project home pages or other useful information. Among other services, this could use the DejaNews service for finding all messages posted by a person. To be used when there is no hard link. Once a home page is found and verified, a hard link will be introduced.

4.8 Usage tracking

The system collects data on its use to aid in evaluation. The system collects data on each session; with the agreement of the user, sessions can be linked to users for cross-session analysis. Some examples of analyses that can be derived from such detailed data are the number of times each page was visited, the number of times each link was followed, the number of times a term was used in searching, the usage of each type of document or link, and, more interestingly, patterns of pathways through the system.

5 Program testing and usability studies

The document processing tools will be tested for functionality using a variety of sample documents drawn from the NIAAA, NIDA, CSAP, and CESAR Web sites and long machine-readable documents obtained from these agencies We will also collaborate with the Thesaurus Linguae Graecae project in testing tools that might be useful for their purposes. Interactive tools will be tested by program staff and volunteer students. Usability tests will call on volunteers associated with the Center for Substance Abuse Research on the College Park Campus and remote volunteers from SALIS (Substance Abuse Librarians and Information Specialists, using methods described in Nielsen1993.

6 Conduct of the project and roles of project personnel

The principal investigator will provide overall guidance and oversight, as well as design for tools that require research and development, using 8 hours/week general research time. The two consultants will be available to give advice or discuss specific problems.

The systems analyst/programmer to be hired for the project will set up the computer environment, install tools that are ready to run, modify others (or write interface modules to link them into the system), and write new tools as needed. This needs to be a person who is very savvy in programming related to the Web and to text-processing applications (needs to know Java, Perl, CGI scripts, Python, CORBA, database interfaces, Z39.50, and tools such as Lotus Notes). Since the result of this project is to be a portable system, it needs the unified effort of one highly knowledgeable person, rather than a combination of theses from several graduate students. The pay for this person is set to be competitive.

The graduate assistants will do searching on the Web and in bibliographic databases to identify tools that exist or materials that can help in developing new tools. They will also write and/or edit program documentation and other project reports, maintain the project Web site, and be involved in functionality and usability testing.

In addition to the regular project personnel, students of the College of Library Information Services will be involved in specific tasks as part of their course work in regular classes (for example, LBSC 794 Software Evaluation or LBSC 795 Human-Computer Interaction) or through independent studies.

References

This bibliography lists references that inspired ideas for our design, support the design, or describe tools or elaborate on individual features of our design. It is arranged according to the proposal outline.

4 The Vision. General references

Furnas, George W. 1998; Raunch, Samuel J. **Considerations for information environments and the NaviQue workspace.** *ACM Conference on Digital Libraries.* 1998: 79-88.

Twidale, M.B. 1997; Nichols, D.M.; Paice, C.D. **Browsing is a collaborative process.** *Information Processing and Management.* 1997.11; 33(6): 761-83.

Cousins, Steve B. **DLITE: A user interface for digital libraries.** PhD thesis, Stanford University, 1997.

Cousins, Steve B. 1997; Paepcke, Andreas; Winograd, Terry; Bier, Eric A.; Pier, Ken. **The Digital Library Integrated Task Environment (DLITE).** *ACM Conference on Digital Libraries.* 1997: 142-51.

Phelps, Thomas A. 1997; Wilensky, Robert. **Multivalent Annotations**. Peters, Carol, ed.; Thanos, Constantino, ed., **Research and advanced technology for digital libraries.** *ECDL '97.* 1997: 287-303.

Wilensky, R. 1998; Phelps, T. **Multivalent documents: from presentation to collaboration.** *DLI 98 Project-wide workshop.* 1998. http://HTTP.CS.Berkeley.EDU/~wilensky/ucb-mvd.ppt

Schmidt, Joachim W. 1997; Schroder, Gerald; Niederee, Claudia; Matthes, Florian. **Linguistic and architectural requirements for personalized digital libraries**. *Int. J. Digit. Libr.* 1997.4; 1(1): 89-104.

Takahashi, Junichi 1998; Kushida, Takayuki; Hong, Jung-Kook; Sugita, Shigeharu; Kurita, Yasuyuki; Rieger, Robert; Martin, Wendy; Gay, Geri; Reeve, John; Loverance, Rowena. **Global digital museum: multimedia information access and creation on the Internet.** *ACM Conference on Digital Libraries.* 1998: 244-53.

Bishop, Ann 1994. **Toward functional requirements for the digital library (based on focus group interviews with faculty and students).** Draft 1994.12.20. http://anshar.grainger.uiuc.edu/dlisoc/DLI.specs.html

Wilensky, Robert 1995. **Toward work-centered digital information services.** Berkeley, CA: UC Berkeley, Computer Science Division; 1995.12.28. 18 p. [General philosophy and overview of the Berkeley Digital Library Project]
http://http.cs.berkeley.edu/~wilensky/computer-special-issue.ps.gz
Download through reference on http://elib.cs.berkeley.edu/papers.html

Balasubramanian, V. 1998; Bashian, Alf. **Document management and Web technologies.** *Communications of the ACM.* 1998.7; 41(7): 97-115.

Crane, Gregory, editor-in-chief. **Perseus Project. An evolving digital library.**
A system with a rich structure of object types and links. Links are untyped, but a semantics of link types is clearly underlying the system. http://www.perseus.tufts.edu/

4.2 Public and private information spaces

Sikkel, Klass 1997. **A group-based authorization model for cooperative systems.** *Proceedings of the Fifth European Conference on Computer Supported Cooperative Work (ECSCW '97).* 1997: 345-60.

Romano, Nicholas C. 1998; Nunamaker, Jay F.; Briggs, Robert O.; Vogel, Douglas R. **Architecture, design, and development of an HTML/JavaScript Web-based group support system**. *JASIS*. 1998.7; 49(7): 649-667.

4.3 Information structure

Oines-Kukkonen, Harri 1998. **What is a link**? *Communications of the ACM.* 1998.7; 41(7): 98.

Thüring, Manfred 1995; Hannemann, Jörg; Haake, Jörg. **Hypermedia and cognition: designing for comprehension**. *Communications of the ACM*. 1995.8; 38(8): 57-66.

Lim, Ee-Peng 1998; Tan, Cheng-Hai; Lim, Boon-Wan; Ng, Wee-Keong; **Querying structured Web resources.** *ACM Conference on Digital Libraries.* 1998: 297-298.

4.4 System architecture

Roscheisen, Martin 1997; Baldonado, Michelle; Chang, Kevin; Gravano, Luis; Ketchpel, Steven; Paepcke, Andreas. **The Stanford InfoBus and its service layers: augmenting the Internet with higher-level information management protocols**. Stanford, CA: Stanford University Digital Library Project; 1997.8.
http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1997-0065

Paepcke, Andreas 1998; Baldonado, Michelle; Chang, Chen-Chuan K.; Cousins, Steve; Garcia-Molina, Hector. **Building the InfoBus: A review of technical choices in the Stanford**

digital library project. Stanford, CA: Stanford University Digital Library Project; 1998.6. http://www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1998-0096.

Trevor, Jonathan 1997; Koch, Thomas; Woetzel, Gerd. **MetaWeb: bringing synchronous groupware to the World Wide Web.** *Proceedings of the Fifth European Conference on Computer Supported Cooperative Work (ECSCW '97).* 1997: 65-80.

4.5 Search and other user-system interaction functions

Thesaurus interface

Johnson, Eric H. 1995; Cochrane, Pauline A. **A hypertextual interface for a searcher's thesaurus**. *ACM Conference on Digital Libraries.* 1995. http://csdl.tamu.edu/DL95/papers/johncoch/johncoch.html

Patil, R. 1996. **The Ontosaurus Ontology Browser**. Unpublished ms. Marina del Rey, CA: USC/Information Sciences Institute; 1996.

Cooper, James W. 1997; Byrd, Roy J. **Lexical navigation: visually prompted query expansion and refinement**. *ACM Conference on Digital Libraries.* 1997: 237-46.

Search functions

Baldonado, Michelle 1997; Wang, Q.; Winograd, Terry. **SenseMaker: an information-exploration interface supporting the contextual evolution of a user's interests**. *CHI '97*. 1997. www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1996-0048

Chen, Hsinchun 1998; Houston, Andrea; Sewll, Robin; Schatz, Bruce R. **Internet browsing and searching: User evaluations of category map and concept space techniques**. *JASIS*. 1998.7; 49(7): 582-603.

Using document structure in searching

Lim, Ee-Peng 1998; Tan, Cheng-Hai; Lim, Boon-Wan; Ng, Wee-Keong. **Querying structured Web resources.** *ACM Conference on Digital Libraries.* 1998*:* 297-8.

Searching multiple databases.

Gravano, L. 1997; Chang, C-C. K.; Garcia-Molina, H. **STARTS: Stanford proposal for Internet meta searching.** *Proc. of the ACM SIGMOD Conference*. 1997: 207-18.

Gravano, Luis 1997; Chang, Kevin; Garcia-Molina, Hector; Lagoze, Carl; Paepcke, Andreas. **STARTS. Stanford Protocol Proposal for Internet Retrieval and Search**. 1997.1. Download PostScript version from http://www-db.stanford.edu/~gravano/starts_home.html

Davidson, S. B. 1997; Overton, C.; Tannen, V.; Wong, L. **BioKleisli: a digital library for biomedical researchers**. *Int. J. Dig. Lib*. 1997.4; 1(1): 36-53. http://thymine.iss.nus.sg:8080/biokleisli.html

Rodgers, R.P. Channing 1995. **Automated retrieval from multiple disparate information sources: The World Wide Web and NLM's Sourcerer project**. *JASIS*. 1995.12; 46(10): 755-764.

Query mapping

Chang, Chen-Chuan K. 1996; Garcia-Molina, Hector; Paepcke, Andreas. **Boolean query mapping across heterogeneous information sources**. *IEEE Transactions on Knowledge and Database Engineering*. 1996.8; 8(4): 515-522. www-diglib.stanford.edu/cgi-bin/WP/get/SIDL-WP-1995-0022

Result mapping and integration

Chang, Cehn-Chuan K. 1998; Garcia-Molina, Hector. **Conjunctive constraint mapping for data translation.** *ACM Conference on Digital Libraries.* 1998: 49-58.

Baldonado, Michelle; Chang;Kevin; Gravano, Luis; Paepcke, Andreas. **The Stanford Digital Library Metadata Architecture.** www-diglib.stanford.edu/diglib/pub/delos.html

Stanford Digital Library Project. **InterBib**. Provides three facilities: conversion of bibliographies among different formats, the processing of documents to include bibliographies, and the collaborative accumulation of bibliographies that can be searched http://www-interbib.Stanford.EDU/~testbed/interbib/interbibInfo.html

Conry, T.J. 1986; Hushon, J.M. **Critical issues in microcomputer gateway design. The Micro-CSIN experience**. *Chemical Information Bulletin*. 1986 Summer; 38(2): 18 [A vintage system that searches several databases for data on chemical substances and integrates the results into a single report.]

Arrangement of search results

Cutting, D. R. 1992; Karger, D. R.; Pederson, J. O.; Tukey, J. W. **Scatter/gather: a cluster-based approach to browsing large document collections**. *Proceedings of ACM/SIGIR.* 1992: 318-29.

Hearst, Marti A. 1995; Karger, David R.; Pederson, Jan O. **Scatter/gather as a tool for the navigation of retrieval results.** *Proceedings of AAAI Fall Symposium on Knowledge Navigation.* 1995.

Hearst, Marti A. 1996; Pederson, Jan O. **Reexamining the cluster hypothesis: Scatter/gather on retrieval results.** *Proceedings of ACM/SIGIR.*. 1996.

Kanada, Yasusi 1998. **Axis-specified search: a fine-grained full-text search method for gathering and structuring excerpts.** *ACM Conference on Digital Libraries.* 1998: 108-17.

Sahami, Mehran 1998; Yusufali, Salim; Baldonado, Michelle Q.W. **SONIA: a service for organizing networked information autonomously.** *ACM Conference on Digital Libraries.* 1998: 200-9.

Lin, Xia 1997. **Map displays for information retrieval.** *JASIS*. 1997.1; 48(1): 40-54.

Enriched document display

Carr, L. 1995; De Roure, D.; Hill, G.; Hall, W. **The distributed link service: a tool for publishers, authors, and readers.** *Proceedings of the Fourth World Wide Web conference.* 1995. http://wwwcosm.ecs.soton.ac.uk/dls/link_service.html

4.6 Reading tools and document creation and editing tools

Streitz, N. et. al 1992. **SEPIA: a cooperative hypermedia authoring environment.** *Proceedings of the ACM Conference on Hypertext*. 1992: 11-22.

Chen, Chaomei 1997. **Writing with collaborative hypertext: analysis and modeling.** *JASIS.* 1997.11; 48(11): 1049-1066.

Reading, annotation, and linking tool

Marshall, Catherine C. 1997. **Annotation: from paper books to the digital library.** *ACM Conference on Digital Libraries.* 1997: 131-40.

Roescheisen, M. 1995; Mogensen, C.; Winograd, T. **Shared Web annotations as a platform for third-party value-added information providers: architecture, protocols, and usage examples.** Stanford Integrated Digital Library Progect, Computer Science Dept., Stanford Univ.: Stanford Univ., November 1994/April 1995. (STAN-CS-TR-97-1582)
(Research prototype called ComMentor no longer maintained)

Alexa, Melina; Rostek, Lothar. **TATOE: Text Analysis Tool with Object Encoding.** http://www.darmstadt.gmd.de/~rostek/tatoe.htm

See also multivalent docs

Authoring tool

Kirby, Andrew 1995; Rodden, Tom. **Contact: support for distributed cooperative writing.** *Proceedings of the Fourth European Conference on Computer-Supported Cooperative Work (ECSCW '95).* 1995: 101-16.

4.7 Document processing and link generation tools

Document segmentation tool

Hearst, Marti A. 1994. **Multi-paragraph segmentation of expository discourse.** Berkeley, CA: UC Berkeley, Computer Science Division; 1994.1.12 (UC Berkeley Computer Science Technical Report No. UCB/CSD-94-790). http://sunsite.berkeley.edu:80/Dienst/UI/2.0/Describe/ncstrl.ucb%2fCSD-94-790?

Hearst, M. 1993. **TextTiling: A quantitative approach to discourse segmentation.** Berkeley, CA: University of Berkeley, 1993. http://sunsite.berkeley.edu:80/Dienst/UI/2.0/Describe/ncstrl.ucb%2fS2K-93-24?

Hearst, M. 1993; Plaunt, C. **Subtopic structuring for full-length document access.** *Proceedings of SIGIR.* 1993: 59-68.

Rus, D. 1995; Summers, K. **Using white space for automated document structuring.** *Advances in digital libraries. Springer-Verlag lecture notes in computer science.* 1995.

Wang, D. 1989; Srihari, S. **Classification of newspaper image blocks using texture analysis.** *Computer Vision, Graphics, and Image Processing.* 1989; 47.

Citation link and citation index tool

Giles, C. Lee 1998; Bollacker, Kurt D.; Lawrence, Steve. **CiteSeer: an automatic citation indexing system.** *ACM Conference on Digital Libraries.* 1998: 89-98.

Bibliographic citation parser

See Giles 1998, just above

Summarization tool

XEROX PARC

Automatic / computer-assisted document categorization and indexing tool

A large body of work exists in this area, going back almost 40 years. Just a sampling of recent work is given here.

Shin, Dongwook 1997; Nam, Sejin; Kim, Munseok. **Hypertext construction using statistical and semantic similarity.** *ACM Conference on Digital Libraries.* 1997: 57-63.

Dolin, R. 1998; Agrawal, D.; El Abbadi, A.; Pearlman J. **Using automated classification for summarizing and selecting heterogeneous information sources.** *D-Lib Magazine*. 1998.1. www.dlib.org/dlib/january98/dolin/01dolin.html

OCLC Office of Research. **The Scorpion Project.** Building of tools for automatic subject recognition based on well known schemes like the Dewey Decimal System. http://orc.rsch.oclc.org:6109/

Thompson, Roger 1997; Shafer, Keith; Vizine-Goetz, Diane. **Evaluating Dewey concepts as a knowledge base for automatic subject assignment.** *ACM Conference on Digital Libraries.* 1997: 37-46.

Fisher, David E. 1994. **Topic characterization of full length texts using direct and indirect term evidence.** Berkeley, CA: UC Berkeley, Computer Science Division; 1994.5 (UC Berkeley Computer Science Technical Report No. UCB/CSD-94-809). http://sunsite.berkeley.edu:80/Dienst/UI/2.0/Describe/ncstrl.ucb%2fCSD-94-809?

Leung, Chi-Hong 1997; Kan, Wing-Kay. **A statistical learning approach to automatic indexing of controlled index terms.** *JASIS*. 1997.1; 48(1): 55-66.

5 Program testing, usability studies, user studies

Nielsen, Jacob. **Usability enginering**. Boston: Academic Press; 1993.