Summary

We propose two broad classes of service to the NSDL. First, we will provide **automatic linking** services that automatically bind key words and phrases to supplementary information. Such automatic linking services are already in place in the Perseus Digital Library. These services will help students, professionals outside a particular discipline, and the interested public to read documents full of unfamiliar technical terms and concepts. Astronomy students and curious amateurs may need to see expansions of some acronyms — e.g., MACHO: massive compact halo object, such as neutron stars and brown dwarfs — or pictures of "Kuiper belt objects." These services can be of particular help to undergraduates as they shift from textbooks to scientific literature: the student struggling with research papers on bioluminescence, for example, will be able to locate information about particular chemical processes or relevant species of echinoderms. Second, we will base automatic linking on **authority control** of names and terms and on links among different authority lists such as thesauri, glossaries, encyclopedias, subject hierarchies, and object catalogues.

For our **Services for a Customizable Authority Linking Environment** (SCALE) to support all levels of reading in the National STEM Digital Library, we propose to

- create and maintain authority lists of technical terms and concepts harvested from OAI metadata, extracted from full text, and imported from existing authority lists
- extend our current automatic hypertext capabilities to embed glosses of technical terms and links to related passages within HTML, PDF, and XML documents in the NSDL
- provide term detection and document linking through SOAP-based web services
- evaluate the interface to and functionality of our service in cooperation with the National Virtual Observatory
- customize the service through explicit user configuration and adaptive learning of preferences
- collect annotations on the quality of links to improve precision and provide a training set for future systems

Much of the work at the Tufts University Perseus Digital Library Project and the Johns Hopkins Digital Knowledge Center — both funded under DLI-2 — has already focused on exploiting various kinds of authority lists (gazetteers, biographical dictionaries, dictionaries, glossaries of technical terms, and name authority files) for the automatic generation of hypertext links [8] and for visualizations such as automatically generated dynamic maps and timelines [30, 31]. Such link generation complements the current practice of automatic identification and aggregation of citations. This proposal sets out to augment and transfer our technology for managing authority lists, converting this from a research effort to an institutionalized service serving a wider community.

The support requested will allow two complementary shifts. First, we will be able to extend and refine our existing software base. While much can be done with the text matching and categorization algorithms, our main focus will be upon developing refined user tools that NSDL participants can use over the Web. Second, we will transfer functionality from its current home in our research labs to Tufts' Digital Collections and Archives (DCA) and Johns Hopkins' Digital Knowledge Center (DKC). This move is a natural one and continues an on-going partnership with shared technology resources: for the past four years, the Tufts University archives (now DCA), Perseus, and the DKC have collaborated closely in various projects and technologies. The Principal Investigator, Gregory Colati, is the Director of the Digital Collections and Archives group that will provide the long-term home for this service. Gregory Crane, head of the Perseus Digital Library Project and PI on the DLI-2 grant "A Digital Library for the Humanities," is co-PI and responsible for overseeing the technical development and transfer of the technology. The Johns Hopkins PI, Sayeed Choudhury, is the Director of the Digital Knowledge Center and PI on the DLI-2 "Digital Workflow Management," and will manage the evaluation of the service and development of adaptive authority linking technology. Alexander Szalay, professor of astronomy at Johns Hopkins and head of the NSF-funded National Virtual Observatory, is co-PI and will involve the NVO in incorporating its knowledge bases for astronomy and evaluating the system for the NVO's audiences from K-12 to higher education.

1 Motivation and Scope

Unfamiliar technical terms pose a well-known barrier to communicating scientific information [14]. Whether the reader is a beginning student or a specialist in another discipline trying to survey a new field, unfamiliar terminology can confuse and, if a superficially familiar term in a new field has a distinct usage, even mislead. Although the National Virtual Observatory, for example, is committed to outreach and education [25], astronomy students and curious amateurs may need to see exemplary pictures of "Kuiper belt objects" or expansions of some acronyms — e.g., MACHO: massive compact halo object, such as neutron stars and brown dwarfs — as they move through this resource. Likewise, undergraduates who are moving beyond the carefully controlled information environment of the textbook will encounter materials that refer to new concepts and, in some cases, old familiar concepts with new names. A student working on bioluminescence, for example, may encounter references to particular chemical processes or species of echinoderms. Students of environmental policy need to ask a broad range of questions from unfamiliar fields: how seismic is the geology of an area? which species are exotic? what is the life cycle of a particular fish?

We propose to augment readers' ability to make connections among scientific information in the National Science Digital Library. Exploiting existing metadata, full text, glossaries, thesauri, and object catalogues, we will extend the abilities of readers from secondary education to professionals to life-long learners. In particular, we aim to encourage undergraduate research — smoothing the transition from absorbing textbooks and problem sets to exploring and creating knowledge.

Our **Services for a Customizable Authority Linking Environment** (SCALE) will provide two basic functions: automatically generating links in and annotations to NSDL documents to aid reading; and, as a basis for this linking, aggregating and managing authority lists of names, technical terms, and concepts.

Supporting reading through **automatically generated links** is a process sometimes called "just-in-time information retrieval" [26, 5, 17, 12]. The Perseus Digital Library System (PDLS) has been providing automatic linking services as part of its production digital library system for more than five years. DLI-2 support has allowed Perseus to confront the problems of customizing automatic linking services for multiple collections. While substantial work remains to be done on customization, the automatic links are, at present, the most heavily used feature in the Perseus DL. The current service supports a substantial user base – the Perseus DL is currently (spring 2002) serving c. 9,000,000 pages per month. We have now established an automatic linking proxy service as an Open Archives Initiative service: readers can call on this service to add links to third-party HTML, plain text, and PDF documents through this service, creating links from matched key words and phrases to glossaries, images, and other supplementary materials. We also have substantial experience in the information extraction strategies that extract key words and phrases from a variety of authority lists, including dictionaries, encyclopedias, glossaries, grammars, textbooks, review articles and similar sources. Originally designed to support documents on Greco-Roman antiquity, the automatic linking service has been adapted to subjects such as early modern English, the history and topography of London, the history of mechanics, and the US Civil War. It now plays a fundamental role in the Dibner Institute's History of Recent Science Project (`http://hrst.mit.edu`). We propose to expand this service as a whole and to augment its customization component for the NSDL.

As in Perseus, automatic linking in documents is based on **authority control** of names and terms and links among different authority lists. While authority lists are already familiar to librarians and many collection developers, we believe that NSDL services that allow authors to choose recognized terms and headers will find a much broader audience in their efforts to make their materials more useful in the distributed NSDL. Researchers at both the Perseus Digital Library Project at Tufts and the Digital Knowledge Center at Johns Hopkins' Sheridan Libraries have spent years developing tools to support this task. The DKC team has particular experience with integrating Library of Congress authority lists [9, 37], while the Perseus team has worked extensively with a range of heterogeneous sources (e.g., pre-existing dictionaries and encyclopedias on various topics). Both groups confront the problem of scalability: in rapidly growing enivironments such as the Perseus DL and especially the NSDL, we need to draw heavily on automated methods and on contributors without professional training as

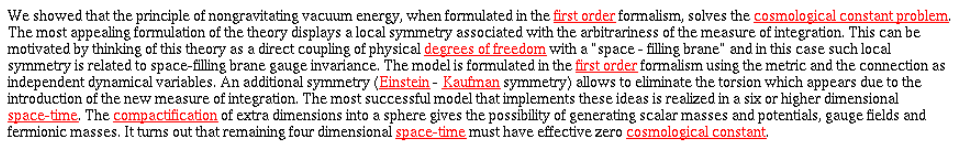


Figure 1: When reading a document in the NSDL, a student can get more information from SCALE about the "cosmological constant," "compactification," or other terms.

cataloguers. We therefore stress probabilistic methods that can automatically identify key words and phrases (e.g., recognize that the "Springfield" in a given passage is in Missouri and not Massachusetts), assess with reasonable accuracy the probability that its identification is correct, display those confidences to users, and allow collection managers to use those confidence levels to target strategic interventions (e.g., show all entries where the confidence level is below or above a given level). We propose to build on this foundation to provide generalized services for the NSDL.

Although automatic linking and authority control can be treated as separate services, their power increases substantially when they are used together. Librarians worked with authority lists for generations before the advent of digital computers, while lists of key words and phrases do already support useful automatic linking without formal authority control [8]. The two services, nevertheless, increase substantially in their power if interconnected, and we believe that a joint proposal that integrates both services will be more effective than separate approaches. Since feedback from some already developing services for the NSDL suggests that service integration will become the biggest general technical challenge for the NSDL and CIS,[1] it makes general sense to develop related services together. And these two services in paricular need to interact with one another: while different reference works may cite "Mark Twain" or "Samuel Clemens", the authority manager can be used to link the two, so that references to "Mark Twain" and "Samuel Clemens" generate links to the same source (and to an explanation that these are alternate references to the same person). We therefore propose this joint project, in which we develop the resources in tandem. The resulting services will become part of the Tufts University Library's Digital Collections and Archives (DCA) group, which provides digital library services to the university as a whole. This transfer from research to production at DCA is a natural one, since the DCA has already integrated the Perseus Digital Library system into its own production environment. The DCA will then provide services and a repository for the resulting software, which will be distributed on an open-source basis.

When students first encounter unrestricted research materials, they may be daunted by unfamiliar terms or concepts. SCALE's proxy service adds links to the document, tailored to the student's knowledge and interests, that provide more information about such terms as "cosmological constant" or "compactification" (figures 1 and 3). A user selecting "cosmological constant" sees links to five glossary entries and 85 articles, in this case harvested from registered OAI data providers. The glossary entry may contain automatically generated links to yet more terms (figure 2).

Similarly, a biology student may seek further information on "dinoflagellates" (a type of organism) or "bioluminescence" (a biochemical process: see figure 4) and then explore the biochemical process in more depth (figure 5).

Scalability is an important goal for any NSDL service. Although introductory texts may be carefully edited to explain terminology on first use, someone navigating a web site may jump into a document in the middle, missing the explanation. It would also be onerous to expect that all materials used for instruction would be manually glossed, and even if they were, at what level ought the terminology be explained? Non-specialists might want all manner of unusual terms glossed and marginal help offered, but specialists could do without these distractions and prefer links to recent papers on related topics.

[1] On March 9, 2002, in a workshop at the Dibner Institute at MIT, Kaye Howe and James Frew reported on Core Integration work in the NSDL. They cited service integration as a focus of current discussions and development.

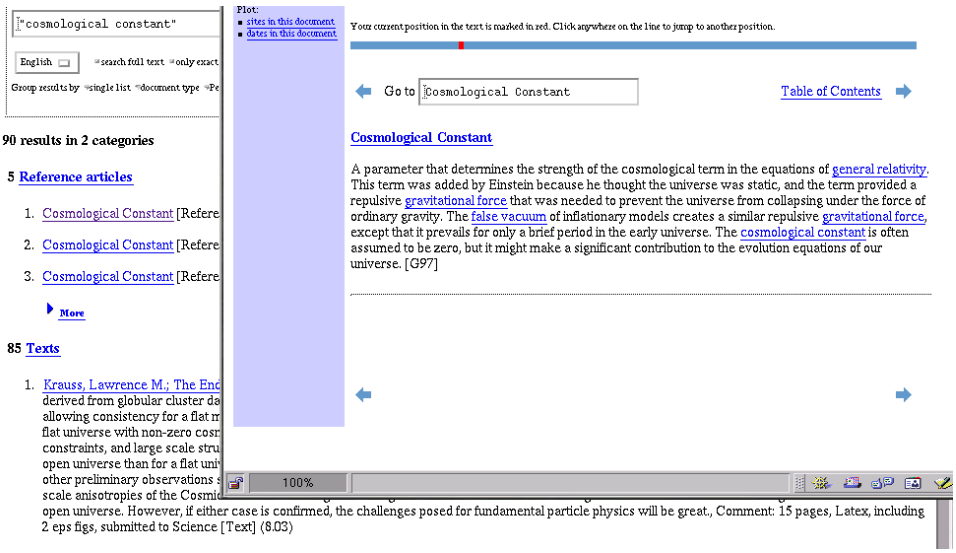


Figure 2: The user has selected the link to "cosmological constant." In addition to five glossary entries, the service finds 85 articles related to that term. The glossary entry contains automatically generated links to yet more terms.

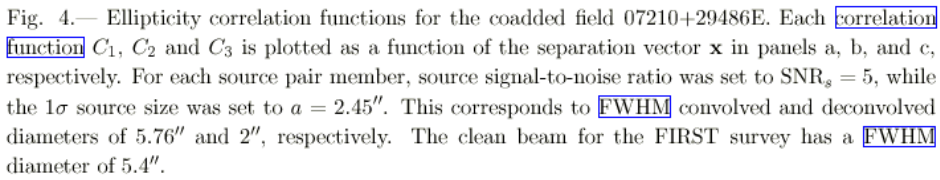


Fig. 4.— Ellipticity correlation functions for the coadded field 07210+29486E. Each correlation function C_1, C_2 and C_3 is plotted as a function of the separation vector $\mathbf{x}$ in panels a, b, and c, respectively. For each source pair member, source signal-to-noise ratio was set to $\mathrm{SNR}_s = 5$, while the 1σ source size was set to $a = 2.45''$. This corresponds to FWHM convolved and deconvolved diameters of $5.76''$ and $2''$, respectively. The clean beam for the FIRST survey has a FWHM diameter of $5.4''$.

Figure 3: The linking service also applies to PDF documents. The acronym is not expanded anywhere in the document. Without such linking the non-specialist will have to hunt for a glossary or explanation.

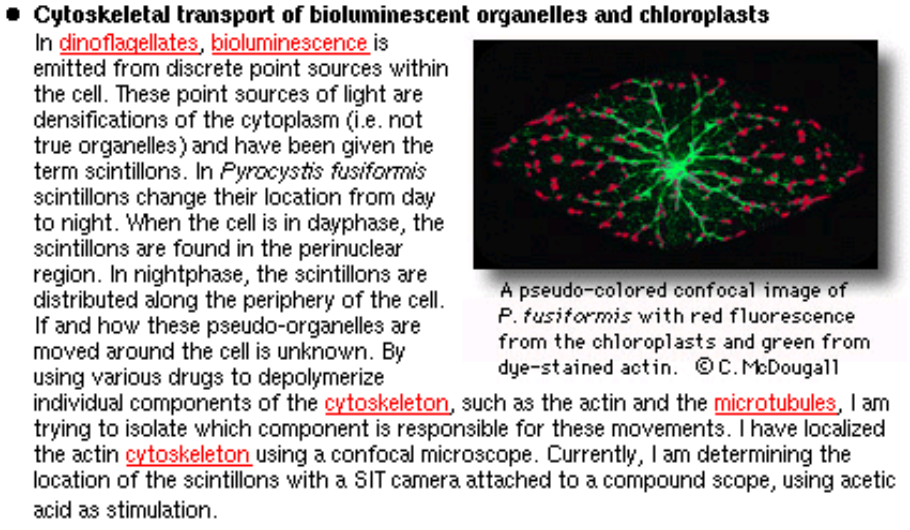
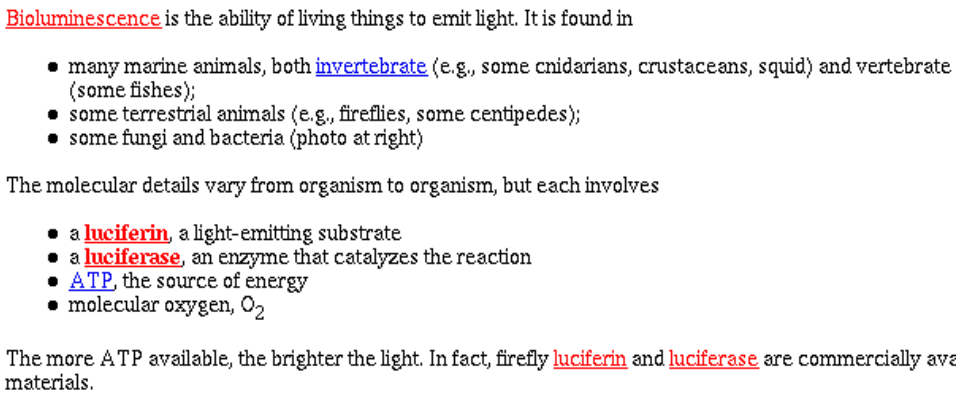


Figure 4: In this biology document, a student can get more information from SCALE about "dinoflagellates," "bioluminescence," or other terms.


Bioluminescence is the ability of living things to emit light. It is found in

- many marine animals, both invertebrate (e.g., some cnidarians, crustaceans, squid) and vertebrate (some fishes);
- some terrestrial animals (e.g., fireflies, some centipedes);
- some fungi and bacteria (photo at right)

The molecular details vary from organism to organism, but each involves

- a **luciferin**, a light-emitting substrate
- a **luciferase**, an enzyme that catalyzes the reaction
- ATP, the source of energy
- molecular oxygen, O_2

The more ATP available, the brighter the light. In fact, firefly luciferin and luciferase are commercially ava
materials.

Figure 5: In this article about bioluminescence, pre-existing links in HTML or PDF can be preserved or overridden. In the illustration above we have used color to distinguish original and added links: blue links (e.g., "invertebrate") were in the original HTML; red links (e.g., "luciferin") were added by the proxy service.

To realize the goal of scalable authority control and linking, we propose to do the following:

- create and maintain authority lists of technical terms and concepts harvested from OAI metadata, extracted from full text, and imported from existing authority lists
- extend our current automatic hypertext capabilities to embed glosses of technical terms and links to related passages within HTML, PDF, and XML documents in the NSDL
- provide term detection and document linking through SOAP-based web services
- evaluate the interface to and functionality of SCALE in cooperation with the National Virtual Observatory
- customize the service through explicit user configuration and adaptive learning of preferences
- collect annotations on the quality of links to improve precision and provide a training set for future systems

Our own site will provide an interface for browsing documents in the digital library augmented with semantic information and visualizations. We will also provide tools for content creators to refine authority lists and to re-incorporate the enriched documents back into their own repositories (figure 6).

Different audiences need to be able to tailor the nature and amount of contextual information they need or, in the case of instructors, how much and what level of contextual information they want their students to have. Incorporating contextual information within the document itself leads to easier navigation and information discovery.

2 Reading Support in Digital Libraries

Much digital library research concentrates on the problem of finding relevant objects. Once the object is found, the user is left to read or analyze the information on his or her own. The OAI protocol, which stands at the core of the NSDL, follows in this tradition, providing a common interface for search and discovery services.

The Perseus Digital Library Project, by contrast, has historically put the problem of reading support at the core of its efforts: what sorts of resources can we generate automatically to help the language student move through a text? how do we automatically create links between allusions to cultural and historical phenomena in an 1838 description of London and explanatory materials for the reader? Since planning began more than

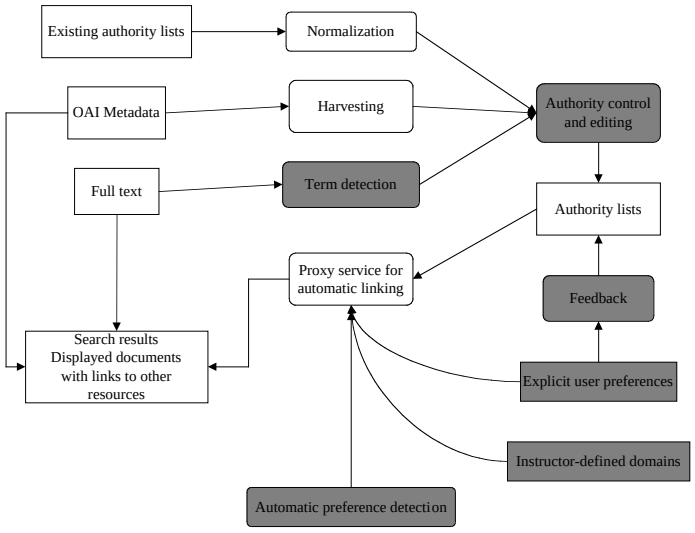


Figure 6: Data and processes in SCALE. White components already exist; grey components will be built under this grant.

fifteen years ago, Perseus has been researching new ways to connect information, using linguistic, citational, and semantic methods [33]. In addition to developing sophisticated linguistic and lexicographic tools to aid the reading of classical Greek and Latin, Perseus has explored the use of reference works and authority lists to create a rich automatic hypertext from digital library materials on the Greco-Roman world. Over the past three years, in collaboration with the Tufts University Archives (now DCA) and with funding from Phase 2 of the Digital Libraries Initiative, Perseus has investigated the process for building a digital library in a new domain: the history and topography of London. By combining narrative histories and fictional works with reference materials, such as the *Dictionary of National Biography* and Wheatley's topographical encyclopedia, and authority lists, such as G.F. Cruchley's London gazetteer and the indices at the backs of books, we have been able to build an automatically hyperlinked collection, with especially rich connections for people, places, and time [8]. Also with NSF funding, Perseus has been collaborating with the Max Planck Institute for the History of Science, in Berlin, building a digital collection on the history of mechanics from Aristotle through Galileo [1]. As part of this effort, we have expanded our linguistic tools to handle Italian, especially the many variant forms that occur in Renaissance texts, and have experimented with tying together technical terms in order to elucidate the mental models in pre-modern science. We have also digitized period reference works such as dictionaries and, most notably, two eighteenth-century scientific encyclopedias by Gehler and Hutton. Finally, we are providing our digital library infrastructure to MIT's Dibner Institute to support a collaborative documentation of recent science.

While Perseus has focused on historical humanities and scientific information, various scientific communities have developed complementary services for their own literature. Starting with the Los Alamos e-print archive in 1991, scientists have turned more and more to disseminating results quickly and cheaply over the Internet. Once materials reached a certain mass, the community began to build and organize services for aggregating and augmenting this virtual scientific database. Among the most notable are CiteSeer, OpenCit, SFX, and the Open Archives Initiative (OAI) [16, 18, 3, 35, 23, 2]. Many services have addressed the issue of distributed resource discovery. To a lesser extent, researchers have focused on activating the citations embedded in most scholarly

documents.

Our proposed service will complement existing citation linking mechanisms and address a range of user needs to augment the reading of scientific information. We can begin this process by mining existing paper and electronic thesauri, controlled vocabularies, and reference works; we will then use this information to annotate documents in the digital library [10, 22, 13]. We can enrich the authority lists by mining information from documents in the NSDL and by providing tools for contributors to the NSDL to specify authoritative information in their own documents.

The NSDL already has two kinds of resources that can be exploited to help readers with terminology. First of all, many of the web materials that are already catalogued have glossary and acronym expansion sections. Among the earth sciences resources in the NSDL-funded Digital Library for Earth System Education (DLESE: `http://www.dlese.org`), there are ten such glossaries. Although most glossaries produced for pedagogical purposes have a structure, the hierarchical organization of many specialized authority lists — such as the National Library of Medicine's Unified Medical Language System (UMLS) or the ACM Computing Classification System — can situate a term in the context of similar terms and in a structured view of a discipline. Terms that co-occur with the term in question can also be automatically extracted from the digital library corpus to provide another view of the way that knowledge is organized [28]. The Perseus team has had good results with calculating word co-occurrences to support working lexicographers [27]. One of the major problems with hand-maintained subject hierarchies and thesauri is the misalignment of terms when moving from domain to domain [29]. Perseus has successfully used information retrieval techniques to cluster similar definitions from one dictionary to another, even when the headwords are in different languages.

While short glosses can be used to inform a reader without too much disruption, manually assigned subject headings and information retrieval techniques can be used to find more discursive documents on a particular topic if more information is needed. The set of subject headings, keywords, titles, authors, and other metadata in the digital library supply terms to be mapped onto documents. In addition to providing just-in-time glossing of concepts, the presence of linked terms in a document will show the topics on which a reader can get more information.

In addition to providing glosses and further information about particular terms, we will also offer suggestions for further reading when a user is browsing documents. Some existing systems use information retrieval methods to find documents most similar to the one the user is examining, usually within a term-document vector space [10, 26]; others approache try to capture the idea of "usefulness" for a particular task [6]. Since users of SCALE will be able to express some preference for the kind of resources they find useful, we will base our recommendations on the pattern of links from the document being read to the other resources in the digital library.

3 Key Word and Phrase Aggregation Services

SCALE needs to handle the hundreds of collections and millions of documents projected for the NSDL. We can achieve the necessary scalability by relying on automatic methods that can benefit from strategic human intervention. The first task is to gather lists of significant names and terms to help us gloss documents for reading and to link to related documents. Our system will extract this information from three sources: the metadata that all NSDL collections will provide through the OAI protocol; structured sources where key terms and phrases are explicitly labelled; and unstructured full text. While we will concentrate on materials formally included within NSDL collections, we will opportunistically harvest glossaries, textbooks, articles, and other resources from beyond the NSDL.

3.1 Basic Metadata Harvesting Services

The infrastructure for the NSDL, as built by the Core Integration group, relies on the Open Archives Initiative's Metadata Harvesting Protocol. This protocol, described in [23] and at `http://www.openarchives.org`,

operates by passing XML-encoded document metadata to harvesting applications in response to queries from a small, pre-defined set. All OAI data providers or repositories must support unqualified Dublin Core, and they may also declare support for community-specific metadata standards. OAI service providers harvest metadata in one or more of these formats so that users and other programs can work with the repositories' resources in aggregate. From its roots in the scientific e-print community, the Open Archives Initiative (OAI) has grown to hold a significant place in digital library interoperability efforts. By lowering the barrier for data providers to expose metadata, the OAI Metadata Harvesting Protocol provides a consistent testbed for new service providers. The OAI has thus successfully positioned itself between high-functionality, complex federation schemes and low-functionality, web crawled search services [2].

Since the initial release of the OAI standard, Perseus has been involved as a data provider. The Perseus Digital Library already exposes all of its document-level metadata via the OAI protocol, and the Perseus team are experimenting with exposing richer metadata for distributed information extraction. Perseus also helped found the Open Language Archives Consortium (OLAC), an OAI community for sharing linguistic information.

Our document management system, the Hopper, also harvests registered OAI data providers to build a testbed for distributed services [32]. From the metadata exposed by these OAI repositories, we extract subject headings, keywords, place names, authors, titles, and other authority information.

We now describe two broad sources for technical terms.

3.2 Tagged Technical Term Aggregation Services

Digitized glossaries, gazetteers, textbooks, encyclopedias, and ontologies inside and outside the NSDL provide a wealth of key words and phrases already identified and explained by subject experts. Also in this category fall protein registries, star catalogues, and the National Virtual Observatory's "concept space" that is used to define relationships between attributes used within its different data collections [25]. The headwords and phrases in these documents can be mined and used very effectively as anchors for automatic linking. Similarly, a growing number of documents contain structured markup in tagsets such as the Text Encoding Initiative [34] and include tags that identify technical terms embedded in full texts. Other documents (such as scientific reviews) may use consistent formatting cues (e.g., bold) to identify technical terms as they are introduced and explained. Such formatting can often be used to extract technical terms in an efficient fashion. If the formatting is ambiguous (e.g., italics are used for technical terms and many other features also) the process is less efficient and term extraction performs no better than on untagged text.

The authority list manager automatically identifies ambiguities — e.g., a term shows up in two glossaries from different fields and with different meanings. The NSDL authority list service will help those creating or maintaining such resources to add alternate formulations of the same concept and to adress similar challenges, but the extracted head words and phrases provide a solid initial foundation. In general, specialized language is designed to reduce ambiguity and the resulting key words and phrases provide much more effective anchors for automatic linking than do normal proper nouns (e.g, "Springfield," "John Smith," "Professor Park").

The Perseus DL has over the past fifteen years mined dozens of print reference works in a variety of subjects for key words and phrases, with recent work exploring domains such as the history and topography of London, early modern English, the American Civil War, the history of mechanics, and the history of recent science (in conjunction with `http://hrst.mit.edu`). Current scientific reference works have proven tractable to the methods that we have developed.

SCALE's authority service will have two basic components. First, we will aggregate a wide range of authority lists already available in the NSDL and other sources, with a particular emphasis initially on astronomy, biology, and earth sciences. Second, we will document how the system works, allowing the producers of new resources to see how they can make their materials interact as efficiently as possible. NSDL participants will be able to run their documents through our system.

3.3 Technical Term Identification Services in Unstructured Text

Most NSDL content will be relatively unstructured HTML, PDF, or plain text. Even documents in XML may encode relatively little structure or fail to mark relevant technical terms and phrases explicitly. Nevertheless, techniques from information retrieval and natural language processing help us detect new terms in running prose. While we have been able to mine millions of useful key words and phrases from structured resources and tagged documents, identifying new technical terms in untagged full text is more challenging and benefits from more human intervention. Nevertheless, the extraction and aggregation of technical terms can be a key element in collection development. We will provide services that help collection developers achieve this goal.

Typically, a library will manage the intellectual quality of the content of its bibliographic records through authority control work. A database management or technical services group would be responsible for identifying irregularities in subject headings, in field use, or in other aspects of the catalog that result in ambiguities and inaccurate search results. This group would then resolve these ambiguities or inaccuracies in the catalog. Some libraries outsource authority control work to professional services such as OCLC/MARS or Library Technologies Inc. These services use complex algorithms to compare subject headings in a catalog with an authorized list, and automatically correct the catalog headings. Typically, they work with a limited authorized list, such as LCSH or MeSH.

Combing thousands of documents for identifiable and authorized terms will result in the discovery of many terms that are just entering the vocabulary of a particular discipline and have not yet been authorized by keyword list or glossary maintainers. Technical terms may be identified automatically by their linguistic properties [20]: they are usually noun phrases, normally without optional modifying adverbs or adjectives [21]. Thus, “acid rain” is a technical term, but “extremely acid rain” is not. Terms may also be identified by statistically significant collocations of words [7, 24] or with a combination of statistical, linguistic, and domain knowledge [19].

Of all linguistic features, noun phrases have attracted attention because of their ability to capture the main topics — the “aboutness” — of expository and academic prose [36, 11, 4]. The mapping of concepts to phrases can be more precise than to words. This precision can not only improve information retrieval but also aid automatic summarization and other applications. Rather than noun phrases in general, many researchers have favored simplex noun phrases (SNPs) for practical reasons. A simplex noun phrase in English can be defined as an optional determiner, followed by zero or more attributive nouns or adjectives, followed by the noun head of the phrase. Not only can SNPs be recognized with a finite state machine on part-of-speech-tagged text, but in English the phrase head, and thus presumably the most important content-bearing word, will be last.

Complex noun phrases, verb phrases, and other structures present greater obstacles to recognition. Also, for indexing and browsing purposes, a complex noun phrase such as *group of children* carries more content in the dependent *children* than in the syntactic head. Much information, however, is still encoded in these more complex structures, especially in genres of text outside modern scientific documents, where strings of attributive nouns and adjectives are quite common. Languages other than English also exhibit structures that are not as simple as English SNPs.

SCALE will report potential new headings through a first-time headings report. A human cataloguer must review each unseen term for accuracy. Some headings on the list will undoubtedly be typographical errors, others will be mis-applied terminology, while still others may be newly-coined terms that have not yet been added to the subject glosses or keyword lists.

New headings determined to be valid — that is, not editorial mistakes or typographical errors — will be posted by SCALE with citation to their first use for authority list maintainers and glossary creators to review and possibly add to their lists. This list could be broken up by academic discipline for ease of review. In this way we will also provide a service to those interested in maintaining and updating their glossaries. These updated glossaries will then be used by SCALE to make further improvements in the quality of links in the documents. It would not be possible or proper for our service to attempt to make decisions concerning the appropriateness of a term in a particular discipline.

Unlike many libraries, however, this service will provide the ability to navigate among several different authority lists. While we will not attempt to create a master list for any field—such as the Getty's Art and Architecture Thesaurus or the UMLS—we will make links among entries in various authorities.

For this proposal, we downloaded and minimally tagged in SGML seven biology glossaries, two astronomical glossaries, four earth science glossaries, as well as the glossary to a USGS biological resources survey. We then exploited these glossaries to create links to technical terms in texts (see above) and used vector similarity information retrieval techniques to create connections among the glossary entries. Thus, we can link the entry for igneous rock in one glossary not only with the entry for igneous rock in another, but also with the definitions for magma, lava, cumulate, porphyritic, glassy, phenocryst, crystal, vesicle, and mafic—all of them types of igneous rocks or materials such as lava pertinent to their formation. These alternative terms can be used to improve recall in searching (by query expansion), to give a reader a wider sense of the semantic field of a term, and to aid switching between different sets of terminologies.

4 Automatic Linking Services

Once metadata are harvested and the key words and phrases are identified, we have services already in place to help users search and browse for documents of interest, or visualize the distribution of places or terms, through the Perseus Digital Library's interface. When they wish to view the content of a resource, they can link to the external site directly, or they can choose to filter documents through a proxy service that provides linking and aggregation services (figures 1, 3, 4, 5, 7). Readers of plain text, HTML, or PDF documents can see links on names and technical terms that link to related resoures. We support different sets of authority lists for different collections. Users could choose to see automatic links from all astronomical or earth science data sources, or could customize the list of authority lists applied (described below).

In addition to this human interface, we will provide an application interface to SCALE. For maximum flexibility, we will use web services protocols such as the Simple Object Access Protocol (SOAP), for transport; and the Web Services Description Language (WSDL), for description of the service's functionality. We expect that once the Core Integration System provides a registry, all of our service descriptions will be fully published there. Existing libraries for Java and Perl ease the creation of SOAP services, and we have taken advantage of them to build test services for term detection and annotation. Web services will allow NSDL content providers or programmers of other services to take advantage of our authority list work.

From the metadata exposed by OAI repositories, we will extract subject headings, keywords, place names, dates, authors, and other authority information. Where the metadata record indicates an address for the resource's content, we will also process that information for terms and names, as well as generating statistics on their uses and on which terms co-occur. We will provide readers with the ability to browse the whole digital library as an automatically generated hypertext: users reading documents filtered through the proxy service on our web site will see unusual terms linked to glosses and more discursive explanatory information. Marginal notes will suggest other resources that are similar to the current document or useful for understanding it; maps with each document will plot the geographic range of the places mentioned. We will at the very least support adding links to documents in HTML and Adobe's Portable Document Format (PDF). We will investigate ways to add links to XML documents in ways that are consistent with various XML style files' support for linking.

At first, in order to guide users to this functionality, we provide a simple searching and browsing interface to find documents of interest (figures 8 and 9). Our interface provides aggregation by static categories such as collection and resource type, but it can also dynamically cluster results by significant phrases.

In the long term, we anticipate exposing rich metadata to the core NSDL system—beyond traditional catalogue-card information—data such as names, dates, and terms extracted from each document and links to related documents. The core integration interface will then be able to harvest our derived metadata and link to our augmented documents.

granite ranges occur in the west, basalt ranges occupy the northwest, rhyolite mountains form the center, and limestone mountains dominate

extreme in the world (Hidy and Klieforth 1990). The Sierra Nevada is primarily responsible for creating this arid, continental climate by capturing moisture from Pacific storm fronts before the moisture reaches the desert (Houghton et al. 1975); similarly, the Rocky Mountains intercept storms from the Gulf of Mexico (Hidy and Klieforth 1990). Local weather patterns are complicated by the mountain ranges that uplift the dispersed moisture, creating mountain storms (Hidy and Klieforth 1990). Thus, precipitation increases with elevation (Billings 1951), and average annual precipitation can be highly variable over small distances.

Figure 7: In addition to glossaries, the service uses gazetteers as an authority list for linking geographic names in texts and map labels.

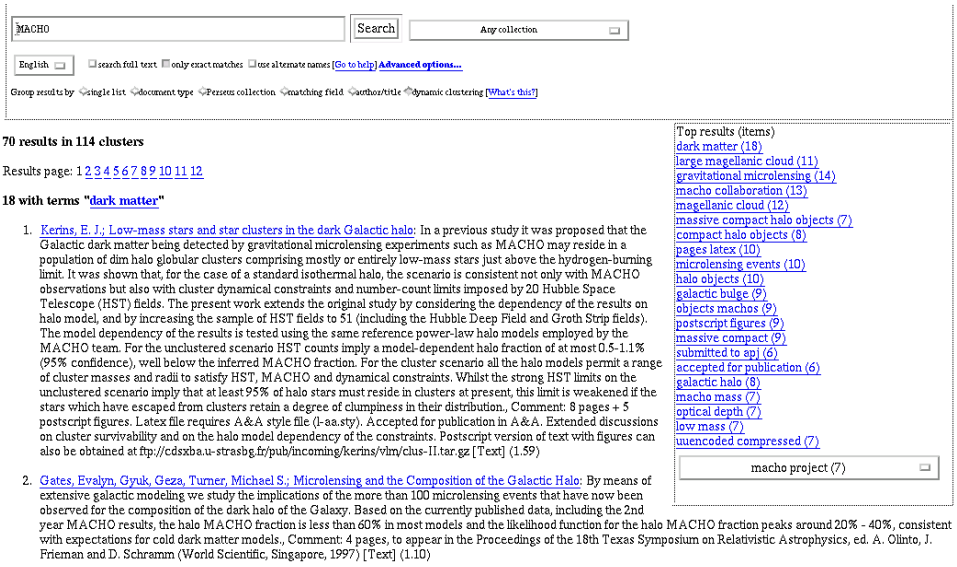


Figure 8: Our interface to harvested data shows how search results for “MACHO” (massive compact halo object) can be clustered by automatically detected phrases that co-occur with that term: “dark matter,” “large magellanic cloud,” “gravitational microlensing,” and so on. Artifacts of the e-print system show up in phrases like “pages latex” and “postscript figure.”

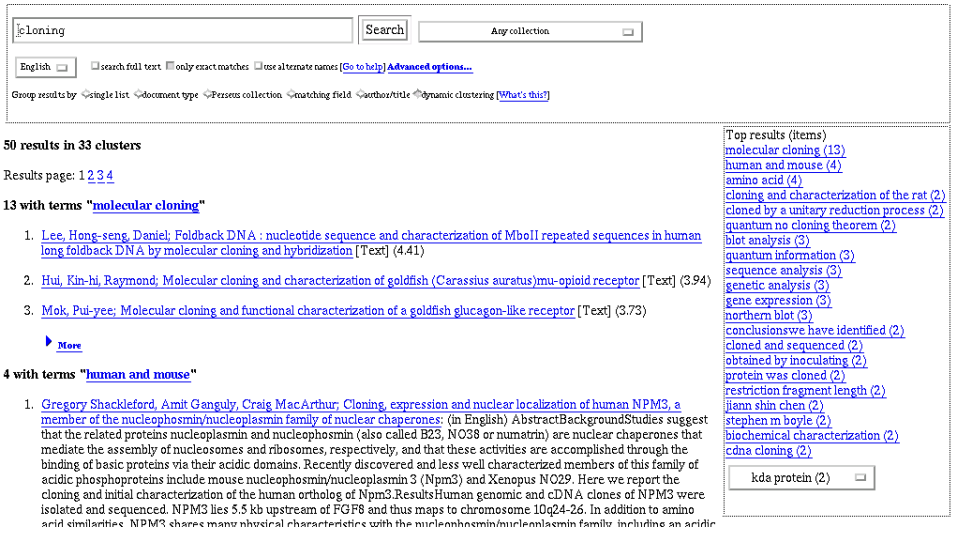


Figure 9: Search results for “cloning” are clustered here by significant phrases such as “molecular cloning,” “human and mouse,” “amino acid,” and so on.

The proxy service also provides the ability to enter arbitrary URLs to have these documents matched (much as some services allow entering a URL to have a document translated). In this case, of course, the link targets will all be documents inside the NSDL, although the link sources will be outside.

The current authority matching system, as implemented in the proxy for human use and the SOAP service for applications, achieves its speed by preselecting a subset of authority lists and then matching on fixed terms. If we can thus determine that a document is about astronomy, we can select an astronomical authority list and ignore potentially confounding terms from biology. While empirical studies show that “one sense per discourse” is a sound linguistic principle [15], there is not always a disciplinary fit between documents and authority lists. We will thus adapt and optimize the DKC’s system for comparing contextual information in the authority data and the target documents to improve precision in authority matching.

5 Customizing Services

Not all users will want to use every sort of linking, suggestion, or visualization. A discipline specialist is unlikely to want simple glosses but may want to see suggestions for further reading or map visualization. A student or a newcomer to the discipline may need new terminology explained and put in context. Consider an undergraduate who, having completed two years of coursework in biology and biochemistry, sets out on his first foray into the research literature. Certain terms and concepts, which his textbooks have highlighted, will be quite familiar. Others will be unfamiliar but central to his topic. Still other terms, while unfamiliar, are incidental to the research at hand. If we wish to help this student make sense of the unrestricted literature, we should direct most effort towards explicating unfamiliar terms that are central to the topic at hand, providing review of known material and the option to branch off into tangential topics.

With this idealized example as motivation, we propose three ways our service can be customized to meet different users’ needs:

1. Users can explicitly choose topics that interest them from an authority list.
2. Users, their instructors, or others can define an ad hoc subdomain that contains documents of interest.

3. The service can examine a user's past and current reading to discern patterns of interest. Thus, the system can identify (1) which technical terms have appeared in recently viewed documents and (2) which new technical terms are particularly prominent in a given document or document collection.

Each form of customization has its own advantages and drawbacks. They can also reinforce each other by being combined. Explicitly choosing topics can provide great flexibility to the user but at a greater expense in time. Adaptive systems can be confusing to some, but they allow users greater flexibility to explore different topics without reconfiguring their reading environments. We expect that customization will form a substantial part of our usability evaluations.

6 Facilities for Feedback and Dissemination

Structured markup standards such as XML promise to bring greater semantic richness to the World Wide Web and digital libraries generally. Technical terms could be tagged as such and linked to persistent vocabulary identifiers; personal and place names could be tagged and linked to biographical dictionaries and gazetteers. Most data on the web, and in many digital libraries, are as yet unstructured, or only structured at the highest level (author, title, section headers, links). The methods for automatic linking and visualization outlined so far rely on relatively simple but effective heuristics and are open to human correction. We have in place a feedback mechanism where a content provider, clicking on a link generated in one of his documents, could rank the most useful resources for the term linked, or correct a wrongly disambiguated name.

Providing authority control for documents in the NSDL will also make it easier for document authors and editors to incorporate rich semantic tagging into their documents. In the Text Encoding Initiative DTD, for example, an author could tag a reference to Springfield not just as `<placeName>Springfield </placeName>` but as `<placeName key='namespace:place:idNum' >Springfield </placeName>`. We will collect the information gained from hand-corrected links and name categorizations from the document authors. We will use this data as a training set for a statistically-based, machine-learning method that advances our heuristic methods of link generation. We plan to advertise the service, gather and compile feedback information, improve our methods of link generation, and then disseminate our findings to the larger community. We will advertise SCALE through both the Tufts University Digital Library, Perseus Digital Library, and NVO web sites. The high traffic of the Perseus site (currently 9,000,000 page views per month not counting embedded images) and of the main Tufts site (15,000 hits per day) ensures that many people will be introduced to the service through serendipitous means. We will also target specific audiences by posting notices on appropriate listservs and websites (the FGDC Clearinghouse weblinks is just one example). Additionally, we will create a brochure for distribution at conferences and workshops such as the ACM/IEEE Joint Conference on Digital Libraries. We also hope to present our findings in the refereed proceedings of these conferences. Finally, we will publish our results in digital library journals such as D-Lib, and through Council on Library and Information Resources and Digital Library Federation publications.

7 Project Organization

Gregory Colati, the Director of Tufts Digital Collections and Archives and the University Archivist, will manage the overall development of the services and their continuation after the end of the grant period. DCA staff will be directly responsible for the human interface to the service and for ongoing authority control.

Sayeed Choudhury, the Director of the Johns Hopkins Digital Knowledge Center and PI on the DLI-2 grant "Digital Workflow Management," will manage the evaluation of the service and the development of adaptive authority linking technology. DKC staff will also work on authority control and on ingesting large pre-existing authority lists. In addition to running the evaluations, they will participate in designing the human interface.

Professor Gregory Crane, head of the Perseus Digital Library Project and PI on the DLI-2 grant "A Digital Library for the Humanitites," is co-PI and responsible for overseeing the technical development and transfer of the technology. The Perseus Project will provide the primary technical skills for the customization and term detection systems and for the application interfaces to the services.

Professor Alexander Szalay will act as co-PI but will not request financial support as the goals of this proposal are consistent with the NSF-funded NVO project. Szalay and the Digital Knowledge Center hold regular discussions regarding the digital library functions of the NVO. For this proposal, Szalay will act as an adviser. This contribution will ensure that the services and interfaces that are developed will be appropriate for the audiences of NVO, ranging from K-12 to higher education. Addtionally, he has agreed to facilitate the process of identifying appropriate subjects for usability testing of interfaces and to act as a liaison to the various outreach partners of NVO.

The DCA and the Perseus Project are in adjacent buildings and interact daily. Tufts and Johns Hopkins will take advantage of their connection to Internet 2 to hold regular videoconferences in the DCA, Perseus, and DKC labs. We also plan one meeting in Baltimore and one in Medford each year to brainstorm development strategies.

8 Implementation Plan

The features and interfaces of SCALE ultimately depend on users' experiences with it in the context of the NSDL. We will thus front-load our development process to bring students and scientists into contact with our service as soon as possible.

8.1 Existing Components

Since, as noted in previous sections, the DCA, DKC, and Perseus already have many working components, we will begin by implementing a test service. Perseus' digital library system, the Hopper, already harvests 434,000 metadata records from registered OAI data providers and integrates them into its searching and browsing tools [32]. The DCA have been using the Hopper for development and production for a year now and have been migrating their collections management system to XML formats. The DKC have been using their authority matching system, based on Bayesian classifiers, as part of their workflow management for capturing the Levy Collection of 30,000 pieces of sheet music [9, 37]. In building a testbed for this proposal, we downloaded several glossaries for astronomy, biology, and earth science to augment the keywords already acquired from harvesting OAI data providers. We also implemented a prototype proxy service that, given an arbitrary URL, adds links from HTML and PDF documents to other resources in the digital library, and rewrites existing links to also pass through the proxy. At present, users of the proxy service can customize the links in a simple way by choosing from a pop-up menu of authority lists. While we will make this proxy service more robust and customizable, we belive that this prototype indicates the feasability and performance of the full service.

8.2 Initial Implementation and Evaluation

The DCA group will begin by harvesting all registered NSDL collections to augment the data providers registered at Cornell. Using the Hopper and the prototype proxy service, the DCA and DKC will make initial modifications to the human interface for the reading environment, concentrating on inline glossing and linking to related documents. The DKC will acquire data from the National Virtual Observatory's 'concept space' to serve as an authority list for astronomical objects, concepts, and processes. This authority list will be used to augment journal articles and other documents archived by the NVO, the NSDL, and other OAI repositories.

The DKC usability specialist will employ a range of methods to evaluate the usability of the services. The usability specialist will conduct a heuristic analysis of the initial interface, in order to identify likely usability and

accessibility problems. The subsequent usability evaluation methods will involve potential users. NVO students, teachers, and scientists, a sample of the target user population of the interface, will be invited to participate in surveys and focus groups to develop interface requirements. Iterative, scenario-based, think-aloud usability tests with NVO users will allow us to observe their interaction with the interface and refine it accordingly. Both the reading environment and the correction interface for subject specialists will be evaluated using these methods. The usability evaluation process will be coordinated with major steps in the development of these services, so that the usability findings will be incorporated into the design. We expect initial evaluation to focus on:

- What aspects of the linking service are favored by practicing scientists, students, or interested amateurs?
- Do students confine themselves to in-line glosses or do links and summaries encourage them to range more widely in the literature?
- Since NVO concepts are keys to enormous data sets as well as scientific articles, do automatic links to (visualizations of) these data sets prove useful to following an author's argument?
- What effects will customized services have on the ease of use of the interface? How can deleterious effects be minimized?

8.3 Continuing Development

Our evaluation of the baseline system will suggest new opportunities for development. As evaluation is going on, however, we expect to work on the four key technologies we have described above:

- improved authority matching based on the DKC system
- automatic detection of new technical terms based on Perseus' work
- customized linking and summarization based on user and document profiling
- tools for organization of, and feedback on, authority lists

The DKC's authority matching system uses Bayesian methods to weight different pieces of evidence for linking to one term or another. We will use other terms, names, and metadata as a feature set to help determine individual term linkages. Perseus' automatic term detection work has had some initial success with the historical scientific documents in the NSF/DFG-funded project Archimedes (`http://archimedes.fas.harvard.edu`). We expect that initial adaptations of these techniques will take three to six months, after which they can be evaluated as part of the on-going usability assessment and retuned. Customization will form the most substantial part of the work for usability evaluation and interface development. This will include end-user customization of the reading environment, as well as tools for the customization of authority lists from terms detected in metadata and full text. Finally, we will incorporate authority lists, as they become available, to enhance the services that we provide and the terms detected automatically. Feedback from authority managers and user preferences will help us refine the authority lists.

We will take advantage of other NSDL services as they go on line. Connecting to the Alexandria Digital Library's service for textual-spatial integration (DUE 0121578), for example, will allow us to use topographic identifiers, rather than possibly ambiguous names, in calculating document similarity and making links.

9 From Research Project to Production Service

A key element in the proposal is to transfer the technology and expertise of the Perseus Project and DKC research and development groups to the Tufts Digital Collections and Archives. Over the course of the grant period, the

DCA will implement the service in a production environment as part of its digital library services. The DCA will develop processes and workflows to manage the service and gather information concerning potential cost-recovery models for supplying the service. Beyond this, the DCA will design and implement the user-interface front end of the system.

After the completion of the grant, the DCA will maintain the service as part of its digital library mission. It is important that this type of activity be moved from research to the front lines of library service and managed by library specialists. The transfer of technology and understanding from research and development projects to library services is a key element in building the digital library. Tufts University is fortunate to have an organizational model specifically designed to support research and development projects and move them into mainstream digital library workflow and management. We expect this service to be used in a number of ways outside the NSDL, all of which support the advancement of research, teaching, and scholarship. Faculty who wish to enhance their students' ability to interrelate documents would be a primary user population. Researchers who want to add value and context to their own writing would also benefit. Most importantly, students of all levels can enhance their learning and develop their research skills by navigating the hypertext of scholarship.

10 Results from Prior NSF Support

This proposal builds on work from the on-going DLI-2 Project, "A Digital Library for the Humanities". In the first three years, we have achieved two primary goals. First, we have created a configurable, scalable digital library system. Second, we have established a set of testbed collections, ranging from antiquity to modern times, in a variety of languages, with distinct audiences and problems. Our goal in years four and five will be to explore the interaction between back-end data structures and front-end features, the use of various human language technologies within a DL, the application of dynamic customization strategies for collections and users; we will seek also to distinguish domain specific from more general DL issues. The work proposed here would be an ideal complement to the research that we will pursue over years four and five of our DLI2 project.

The authority control work in this proposal builds on research by the DKC on the Lester S. Levy Digitized Collection of Sheet Music. The DLI-2 phase of the Levy Project focuses on the development of a workflow management system that will reduce the amount of human labor (and consequently cost) associated with large-scale digitization. Additionally, the project includes the creation of a suite of research tools to facilitate access to digital collections. The cornerstones of the workflow management system include the optical music recognition (OMR) system and an automated name authority control tool (ANAC). The OMR software generates a logical representation of the score for sound generation, music searching, and musicological research. This DLI-2 research has provided the foundation for current work in generalized symbol recognition.

The DKC's automated name authority control tools are designed to disambiguate and match creator names from the Levy Collection to names in the Library of Congress Authority File (LCAF). The tools use Bayesian probability techniques to create a link between a name and an LCAF record and to calculate a confidence measure for that link. Matching is based on the combined probabilities that individual pieces of evidence would yield a correct match. Evidence can be thought of as a weighted set of assertions, with the weights varying according to the strength of the evidence. Our Bayesian approach allows independent weighting of each piece of evidence and each of the possible relationships among first name, last name, and title in the calculation of a final probability. Currently, name (first, middle, last), title (e.g., "Mr.," "Prof.," "Dr."), suffix (e.g., "Esq."), commonness of name, presence of musical terms in LCAF notes fields, and publication (from metadata) after birthdate (from LCAF) are used as evidence. Information from the matching authority entry or entries can be used to link a name or term to other associated information.

References

[1] Alison Abbott. Digital history. *Nature*, 409:556–557, 1 February 2001.

[2] William Y. Arms, Diane Hillmann, Carl Lagoze, Dean Krafft, Richard Marisa, John Saylor, Carol Terrizzi, and Herbert Van de Sompel. A spectrum of interoperability: The Site for Science prototype of the NSDL. *D-Lib Magazine*, 8(1), January 2002.

[3] Donna Bergmark and Carl Lagoze. An architecture for automatic reference linking. In *Proceedings of ECDL 2001*, pages 115–126, Darmstadt, 4-9 September 2001.

[4] Didier Bourigault. Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of the 15th International Conference on Computational Linguistics*, pages 977–981, Nantes, France, 1992.

[5] Jay Budzik and Kristian Hammond. Watson: Anticipating and contextualizing information needs. In *Proceedings of the 62nd Annual Meeting of the American Society for Information Science*, volume 36, pages 727–740, Washington, D.C., 1999.

[6] Jay Budzik, Kristian J. Hammond, Larry Birnbaum, and Marko Krema. Beyond similarity. In *Proceedings of the 2000 Workshop on Artificial Intelligence and Web Search*. AAAI Press, 2000.

[7] Edith Cohen, Mayur Datar, Shinji Fujiwara, Aristides Gionis, Piotr Indyk, Rajeev Motwani, Jeffrey D. Ullman, and Cheng Yang. Finding interesting associations without support pruning. *IEEE Transactions on Knowledge and Data Engineering*, 13(1):64–78, January/February 2001.

[8] Gregory Crane, David A. Smith, and Clifford E. Wulfman. Building a hypertextual digital library in the humanities: A case study on London. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 426–434, Roanoke, VA, 24-28 June 2001.

[9] Tim DiLauro, G. Sayeed Choudhury, Mark Patton, James W. Warner, and Elizabeth W. Brown. Automated name authority control and enhanced searching in the Levy Collection. *D-Lib Magazine*, 7(4), April 2001. `http://www.dlib.org/dlib/april01/dilauro/04dilauro.html`.

[10] Samhaa R. El-Beltagy, Wendy Hall, David De Roure, and Leslie Carr. Linking in context. In *Proceedings of the 12th ACM Conference on Hypertext and Hypermedia*, pages 151–160, Århus, Denmark, August 2001.

[11] David A. Evans and Chengxiang Zhai. Noun-phrase analysis in unrestricted text for information retrieval. In *Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics*, pages 17–24, Santa Cruz, CA, 1996.

[12] H. P. Frei and D. Stieger. Making use of hypertext links when retrieving information. In *Proceedings of the ACM European Conference on Hypertext*, pages 102–111, Milan, Italy, 1992.

[13] Atsushi Fujii and Tetsuya Ishikawa. Utilizing the World Wide Web as an encyclopedia: Extracting term descriptions from semi-structured texts. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 488–495, 2000.

[14] Robert Gaizauskas, Patrick Herring, Michael Oakes, Michelline Beaulieu, Peter Willett, Helene Fowkes, and Anna Jonsson. Intelligent access to text: Integrating information extraction technology into text browsers. In *Proceedings of HLT*, San Diego, CA, 2001.

[15] William Gale, Kenneth Church, and David Yarowsky. One sense per discourse. In *Proceedings of the 4th DARPA Speech and Natural Language Workshop*, pages 233–237, 1992.

[16] C. L. Giles, K. Bollacker, and S. Lawrence. CiteSeer: An automatic citation indexing system. In *Proceedings of the 3rd ACM Conference on Digital Libraries*, pages 89–98, Pittsburgh, PA, 1998.

[17] Gene Golovchinsky. What the query told the link: The integration of hypertext and information retrieval. In *Proceedings of the Eighth ACM Conference on Hypertext*, pages 67–74, Southampton, United Kingdom, April 1997.

[18] Steve Hitchcock, Les Carr, Zhuoan Jiao, Donna Bergmark, Wendy Hall, Carl Lagoze, and Stevan Harnad. Developing services for open eprint archives: Globalisation, integration and the impact of links. In *Proceedings of the Fifth ACM Conference on Digital Libraries*, pages 143–151, San Antonio, Texas, June 2000.

[19] Anette Hulth, Jussi Karlgren, Anna Jonsson, Henrik Boström, and Lars Asker. Automatic keyword extraction using domain knowledge. In *Computational Linguistics and Intelligent Text Processing*, volume 2004 of *LNCS*, pages 472–482. Springer, 2001.

[20] Christian Jacquemin. *Spotting and Discovering Terms through Natural Language Processing*. MIT Press, Cambridge, 2001.

[21] John S. Justeson and Slava M. Katz. Technical terminology: Some linguistic properties and an algorithm for identification in text. *Natural Lanuage Engineering*, 1(1):9–27, 1995.

[22] Hermann Kaindl and Stefan Kramer. Semiautomatic generation of glossary links: A practical solution. In *Hypertext '99: Returning to Our Diverse Roots*, pages 3–12, 1999.

[23] Carl Lagoze and Herbert Van de Sompel. The Open Archives Initiative: Building a low-barrier interoperability framework. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 54–62, 2001.

[24] Caroline Lyon, James Malcolm, and Bob Dickerson. Detecting short passages of similar text in large document collections. In Lillian Lee and Donna Harman, editors, *Proceedings of the 2001 Conference on Empirical Methods in Natural Language Processing*, pages 118–125, 2001.

[25] Paul Messina and Alex Szalay. Building the framework for the National Virtual Observatory. Technical report, California Institute of Technology and Johns Hopkins University, 2001. `http://www.us-vo.org/nvo-proj.html`.

[26] Bradley James Rhodes. *Just-In-Time Information Retrieval*. PhD thesis, Massachusetts Institute of Technology, 2000.

[27] Jeffrey A. Rydberg-Cox. Word co-occurrence and lexical acquisition in Ancient Greek texts. *Literary and Linguistic Computing*, 15(2):121–129, 2000.

[28] Bruce Schatz, William Mischo, Timothy Cole, Ann Bishop, Susan Harum, Eric Johnson, Laura Neumann, Hsinchun Chen, and Dorbin Ng. Federated search of scientific literature. *Computer*, 32(2):51–59, February 1999.

[29] Bruce R. Schatz, Eric H. Johnson, and Pauline A. Cochrane. Interactive term suggestion for users of digital libraries: Using subject thesauri and co-occurrence lists for information retrieval. In *Proceedings of the First ACM Conference on Digital Libraries*, pages 126–133, Bethesda, Maryland, 1996.

[30] David A. Smith. Detecting events with date and place information in unstructured text. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 191–196, Portland, OR, July 2002.

[31] David A. Smith and Gregory Crane. Disambiguating geographic names in a historical digital library. In *Proceedings of ECDL*, pages 127–136, Darmstadt, 4-9 September 2001.

[32] David A. Smith, Anne Mahoney, and Gregory Crane. Integrating harvesting into digital library content. In *Proceedings of the 2nd ACM+IEEE Joint Conference on Digital Libraries*, pages 183–184, Portland, OR, July 2002.

[33] David A. Smith, Jeffrey A. Rydberg-Cox, and Gregory R. Crane. The Perseus Project: A digital library for the humanities. *Literary and Linguistic Computing*, 15(1):15–25, 2000.

[34] C. M. Sperberg-McQueen and Lou Burnard, editors. *Guidelines for Electronic Text Encoding and Interchange*. Text Encoding Initiative, March 2002. Available at `http://www.tei-c.org/P4X/`.

[35] Herbert Van de Sompel and Patrick Hochstenbach. Reference linking in a hybrid library environment, part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment. *D-Lib Magazine*, 5(10), October 1999. See `http://www.dlib.org/dlib/october99/van_de_sompel/10van_de_sompel.html`.

[36] Nina Wacholder, David K. Evans, and Judith L. Klavans. Automatic identification and organization of index terms for interactive browsing. In *Proceedings of the First ACM + IEEE Joint Conference on Digital Libraries*, pages 126–134, Roanoke, VA, 24-28 June 2001.

[37] James W. Warner and Elizabeth W. Brown. Automated name authority control. In *Proceedings of the First ACM+IEEE Joint Conference on Digital Libraries*, pages 21–22, Roanoke, VA, June 2001.