

Developing and Testing the NITLE Semantic Engine (NSE)

A Proposal to

The Andrew W. Mellon Foundation

Clara Yu

The National Institute for Technology and Liberal Education

July 10, 2003

Contents

I.	Executive Summary	4
II.	Introduction	5
	A. The National Institute for Technology and Liberal Education (NITLE) and its Research and Development Mission.....	5
	B. The Problem of Unstructured Data	5
	C. NITLE's Past Work on Semantic Indexing	6
III.	Proposed Technologies.....	7
	A. Further Development and Refinement of Search and Clustering Algorithms	7
	1. Latent Semantic Indexing.....	8
	2. Contextual Network Search.....	9
	3. Clustering and Categorization	10
	B. Experiments on Distributed Processing and Searching across Multiple Databases	12
	1. Distributed Processing	12
	2. Peer-to-Peer Search.....	12
	C. Developing New Technologies and Tools.....	13
	1. Data Visualization.....	13
	2. Information Management.....	13
IV.	Implementation in Research Domains	13
	A. The Humanities	13
	1. History.....	14
	2. Literature	14
	3. Image Database with Textual Metadata.....	15
	B. Bioinformatics.....	15
	C. Weblogs	17
V.	Deliverables	18
	A. Search and Clustering Algorithms	18
	B. Information Management Tools and User Interface Features	18
	1. Auto-Categorization and Archive Management Component.....	18
	2. Peer-to-Peer Searching.....	19
	3. Visual Data Models.....	19
	4. Graphical User Interface (GUI).....	19
	5. Stand-Alone Desktop Application.....	19
	C. Domain-Specific Tools and Features	20
	1. The Literary Toolkit.....	20
	2. Visual Navigation and Relevance Feedback	20
	3. Blog Census.....	20
	4. Bioinformatics	20
VI.	Project Team.....	21
VII.	Assessment and Timeline.....	22

VIII. Proposed Budget	23
A. Project Staff.....	23
B. Demonstration and Dissemination	24
C. Hardware and Storage	24
D. Assessment.....	25
E. NITLE Cost-Share.....	25
IX. Conclusion	25
X. Appendices	26
A. Project Team Curricula Vitae	26
1. Maciej Ceglowski	26
2. John Cuadrado	29
3. Clara Yu	34
B. Curricula Vitae of Evaluators	41
1. Michael Keeble Buckland	41
2. Robert A. McCaughey	46
3. John M. Unsworth.....	49

I. Executive Summary

“Maybe the major defining trend of our time is the proliferation of unstructured information, and the breakdown of tools for managing unstructured information.”

— Barney Pell (of *MetaGame Project* and *Whizbang*)

During the past two years, a team of computer scientists and programmers at the National Institute for Technology and Liberal Education (NITLE) has developed a prototype of the NITLE Semantic Engine (NSE). This prototype was designed specifically to address the universal problem of accessing and organizing large amounts of unstructured digital text. Using mathematical algorithms to index the latent semantic content of documents, the prototype engine has been demonstrated to drastically reduce, if not eliminate, the need for expensive and time-consuming metadata tagging, and to produce results superior to keyword searches in limited test domains.

The goal of the proposed project, for which we request support from The Foundation in the amount of \$596,074, is to develop and test the NSE fully in the following areas: large and varied datasets, contextual networks and probabilistic algorithms that scale gracefully, distributed processing and peer-to-peer search, information visualization and management.

As a search engine, the NSE will allow users to search a data space, which may consist of multiple distributed databases resident at different locations, and find documents that are semantically related to the search criteria, even when there is no direct match to the query. As a categorization tool, the NSE will provide users with a starting point — a computer-generated grouping of the data by content, along with summaries of each cluster — to help users quickly determine its contents.

The system will also provide users with relevance feedback features that let them explore and refine their search results by interacting with the system. An “archivist’s interface” combines the above-mentioned features with visualization of data and support for multiple user views. Taken as a whole, the NSE’s search features will remove many of the rote aspects of metadata creation, while its integrated toolset will greatly facilitate the exploration and analysis of information and the harvesting of knowledge from aggregated data.

The data spheres with which we wish to experiment include such diverse domains as the humanities (e.g., history and literature), science (e.g., bioinformatics), and social software-enabled content (e.g., weblogs). We will work with disciplinary experts in colleges and universities to test the NSE and disseminate the system by holding presentations and demonstrations at national conferences and on college and university campuses.

II. Introduction

A. The National Institute for Technology and Liberal Education (NITLE) and its Research and Development Mission

The National Institute for Technology and Liberal Education was created in 2001 with generous support from The Foundation. It serves as a catalyst for innovation and collaboration for the national liberal arts colleges with which The Foundation works, and assists these colleges in using technology to enhance teaching, learning, scholarship, and information management.

The National Institute, itself an innovation, is an experimental model of networked virtual organization in U.S. higher education. One of its core missions is to ensure that NITLE colleges are able to be actively engaged with the exploration and creation of knowledge.

These colleges constitute a 10 billion dollar industry, with nearly identical operations, clientele, and product. Without a “Research and Development” effort, the entire sector runs the risk of being left behind the times, especially in today’s fast-moving technology environment. An important function of NITLE is to serve as the “Research and Development arm” for liberal arts colleges.

The NITLE Semantic Engine (NSE) is at the center of NITLE’s knowledge R&D mission. It is designed to enable scholars and educators to manage the already overwhelming and ever-increasing volume of data that we encounter in every field of inquiry. NITLE plans to develop a set of tools that will enable researchers to quickly search through large datasets that may be resident in different databases, to interact with the engine to refine the search, and to contribute their knowledge to the collection. We also want to devise a set of visualization and archiving tools for the researcher to use, to facilitate the organization and dissemination of the search results.

B. The Problem of Unstructured Data

Online information is increasing at a truly staggering rate — a pace of growth so rapid that even efforts to estimate the size of the problem have lagged behind. It is estimated that the amount of new print material alone exceeds 24 terabytes/year¹; it is safe to say that the amount of electronic data generated is orders of magnitude greater. In the summer of 2002, the Google search engine counted just over two billion Web pages in its index. A year later, that number has grown to three billion. Even Google engineers confess that their search tools can only reach a small fraction of the Web. And the Web pales in size next to some of the archives and scientific datasets that are now becoming available online².

¹ “How Much Information?”, U.C. Berkeley study, 1998.

² Silverstein, Craig. “How Google Grows.” Conference presentation, O’Reilly Emerging Technology conference, 2003.

By and large, information storage technologies have managed to keep pace with this fantastic rate of growth. Faster and smaller hard drives mean that even a desktop PC can store a library's worth of data. But our ability to search electronic data, key to making sense out of this flood of material, has not substantially improved for the past 20 years.

The shortcomings of keyword search, the current search standard, are clear to anyone who has had to rely on it for more than simple queries. Keyword search suffers simultaneously from being too specific and too general. Lacking any information about meaning, it can only look at literal text, overwhelming the user with tangential results while overlooking potentially relevant documents that don't contain an exact match.

Traditionally, content providers have used metadata tagging to improve the performance of keyword search engines, and allow users to search document collections on a less literal level. Rich metadata makes collections vastly more searchable, and permits the kind of high-level organization that human beings need to be able to effectively navigate large document collections.

Unfortunately, metadata is a very expensive resource, sometimes requiring more time to generate than does the content it describes. Moreover, metadata is rather brittle. Standards and perceptions of what is useful can change over time, requiring an expensive re-indexing of the entire data collection. Rapidly changing collections require considerable metadata generation overhead, and because searches rely on the accuracy of the metadata, documents misclassified in a metadata scheme can become essentially irretrievable.

The weaknesses of keyword searching and the limitations of metadata have prompted extensive research into techniques for extracting information from the actual content of the data being indexed. These efforts have run the gamut from attempts to teach computer programs to understand human language, to pragmatic applications of statistical analysis, to minor features of document structure. In the following, we will describe some of the content-based approaches, and NITLE's adaptation of these "pure research" results in developing an effective information management toolset for the wider knowledge community.

C. NITLE's Past Work on Semantic Indexing

NITLE's initial interest in content-driven indexing grew out of our work with Middlebury College's Center for Educational Technology (CET), one of NITLE's three regional technology centers. CET created a multimedia database, a collection of resources (primarily images) that had been assembled for language teachers who wanted a source of copyright-free material to use in language teaching. The database relied on contributions from individual participants, who were expected to provide descriptions as well as fill in a large number of metadata fields for each media item. We found that while contributors were glad to include detailed English-language descriptions of the images, they were reluctant to fill in metadata fields. Furthermore, faculty users of the database

would search for their interest areas, more often than not missing the metadata categories in the database.

Given this pattern of use, we wondered whether there might be ways to use the full language descriptions themselves, substituting detailed narrative text for the underused metadata descriptors. This question eventually led us to a technique called latent semantic indexing (LSI) that had remained in the research community despite its proven utility and promise in applications.

After initial experiments with the image database in early 2002 yielded encouraging results, we decided to test our LSI implementation on a larger dataset by downloading and indexing a test collection of several thousand news wire stories. The resulting search engine convinced us that LSI was an effective way to sidestep the issue of metadata coding.

We presented our initial results to The Mellon Foundation in May 2002, and subsequently undertook a larger feasibility study, this time indexing a sample journal from the JSTOR archive. To handle the considerable number of documents in this data collection, we devised a technique for subdividing the collection into smaller pieces and searching them as an aggregate. This distributed prototype gave us the first glimpse of the possibility of a modular search collection, where scholars would be able to mix and match individual collections to create a custom search domain.

The second opportunity to test our LSI implementation came in the fall of 2002, when we collaborated with the University of Virginia by providing semantic search capabilities for their *Valley of the Shadow* newspaper archive. This search engine was used extensively by history students doing research using the *Valley* archive, and by all accounts increased the accessibility and usefulness of a collection that previously had been difficult to navigate. The trial also provided us with important data on the types of queries we could expect to see “in the wild.”

Concurrently with the UVA trial, we undertook our first experiments in applying LSI to non-text collections, first with mass spectra of organic compounds, and then with protein data from the Protein Data Bank (PDB). These trials also saw the first use of our clustering algorithms in conjunction with semantic search, to automatically group data into homogeneous categories. Both trials were successful; the results from our protein study were presented at the O’Reilly Bioinformatics Conference (February 2003) to considerable interest.

III. Proposed Technologies

A. Further Development and Refinement of Search and Clustering Algorithms

The diverse algorithms underlying the NITLE Semantic Engine share one common feature: they use statistical patterns in content to make inferences about document similarity. The data model used in the NSE is sometimes called the “bag of words”

approach, since it treats documents as an unordered list of semantic tokens (such as nouns and noun phrases) extracted from the original text. It so happens that patterns of co-occurrence among these tokens reveal important information about their relative conceptual distance. By counting how many times different words occur together across a large document collection, the NSE is able to make good guesses about which words and documents are related, without ever understanding what any of them mean.

This lack of understanding is actually a powerful feature, since it makes the NSE a language-agnostic tool. Given a sufficient number of documents, along with some simple guidance on how to find word boundaries, the engine can return high quality search results for any language, without human intervention. The NSE's clustering algorithms, which rely on measures of distance provided by the initial statistical analysis, are similarly flexible.

The key search algorithms used in the NSE are latent semantic indexing (LSI) and contextual network search (CNS).

1. Latent Semantic Indexing

Latent semantic indexing (LSI; also known as "latent semantic analysis," LSA) is the application to text collections of a well-studied noise reduction technique. LSI uses a vector-space model that represents each document in a collection as a vector in a high-dimensional term space. The key to LSI is a dimensionality reduction step that projects these vectors onto a much smaller set of dimensions, in a process analogous to casting shadow puppets on a wall. This projection, accomplished through a linear algebra technique called *singular value decomposition* (SVD), maps documents onto a set of underlying "semantic dimensions," reducing synonymy and creating overlap between related terms. Documents that share many words in common end up near each other in the reduced space, while dissimilar documents stay far apart.

While the mathematics of singular value decomposition has been known for over a century, the technique as applied to text collections is relatively recent. It was first described in a 1981 Ph.D. dissertation³, and then rediscovered in a seminal 1989 paper⁴ that described its first implementation in a text search engine.

LSI has since been the subject of intensive study in the research community. It is a proven technique that can provide up to a fourfold improvement in recall with minimal loss of precision⁵.

³ Preece, Scott. "A spreading activation network model for information retrieval." Ph.D. thesis, CS Dept., Univ. of Illinois, Urbana, IL, 1981.

⁴ Deerwester, S., Dumais, S., Furnas, G., Landauer, T., and Harshman, R. "Indexing by latent semantic analysis." *Journal of the American Society for Information Science*, 1990, pp. 391-407.

⁵ Ibid.

Offsetting these strengths, however, are several weaknesses. The SVD algorithm does not scale well to large collections, requiring significant computational resources during the indexing phase. Moreover, it is not possible to update or remove documents from the collection without periodically having to perform the expensive re-indexing step⁶. These limitations make LSI less useful for very large or rapidly changing collections.

The opacity of the underlying representation means that LSI is a black-box technique whose parameters must be hand-tuned based on empirical testing. The wrong choice of parameters (such as the number of dimensions in the reduced space) can have a very detrimental impact on performance, and the optimum parameters for LSI will vary from collection to collection⁷.

A final obstacle is the fact that LSI has been patented for search in text collections⁸. This limits its utility to non-text collections (such as protein conformation data), where there are no restrictions on the algorithm's use.

2. Contextual Network Search

Contextual network search (CNS) is NITLE's own coinage for an alternative to LSI that we began to develop in 2003, a variant of a technique called "spreading activation search." This core algorithm was first described in a doctoral dissertation⁹ but, to our knowledge, has never been implemented on a sizable text collection. The search algorithm draws on the same data model as LSI — a term/document lookup matrix — but uses an alternate interpretation that is much easier to understand.

Whereas LSI represents documents as vectors in a term space, CNS represents both terms and documents as nodes in a graph, connected by weighted links. Each link connecting a term and document node represents an occurrence of that term in that document. We can generate a word list for any document by looking at its nearest neighbor nodes; similarly, we can find every document in which a given word occurs by looking at its own nearest neighbors. The simplest scheme to "weight" a link is to assign the number of times the given term occurs in the given document. A simplified sample contextual network for a set of seven "documents" is shown in Figure 1.

When a query comes in (for our purposes here, let's assume it is a single word, such as "ice"), it triggers the graph traversal process, which looks up that word

⁶ Simon, H., and Zha, H. "On Updating Problems in Latent Semantic Indexing." *Technical Report No. CSE-97-011*. Department of Computer Science and Engineering, Pennsylvania State University, 1997.

⁷ Deerwester, S., Dumais, S., Furnas, G., Landauer, T. and Harshman, R., "Indexing by latent semantic analysis." *Journal of the American Society for Information Science*, 1990, pp. 391-407.

⁸ "Computer information retrieval using latent semantic structure". U. S. Patent No. 4,839,853, June 13, 1989.

⁹ Preece, Scott. "A spreading activation network model for information retrieval." Ph.D. thesis, CS Dept., Univ. of Illinois, Urbana, IL, 1981.

and travels along the links that lead from that word to all the documents that contain the word. This list of these “relevant documents” is the search result that would occur if we only did keyword searches.

However, the CNS takes the search a step further by traversing the graph from each of the documents in the list of “relevant documents.” This then leads to new term nodes. In turn we continue the traversal from these new term nodes and visit the document nodes to which these terms are linked. Collecting and sorting the nodes that accumulate weight values beyond a given threshold gives us a ranked result set that we can display on a search page.

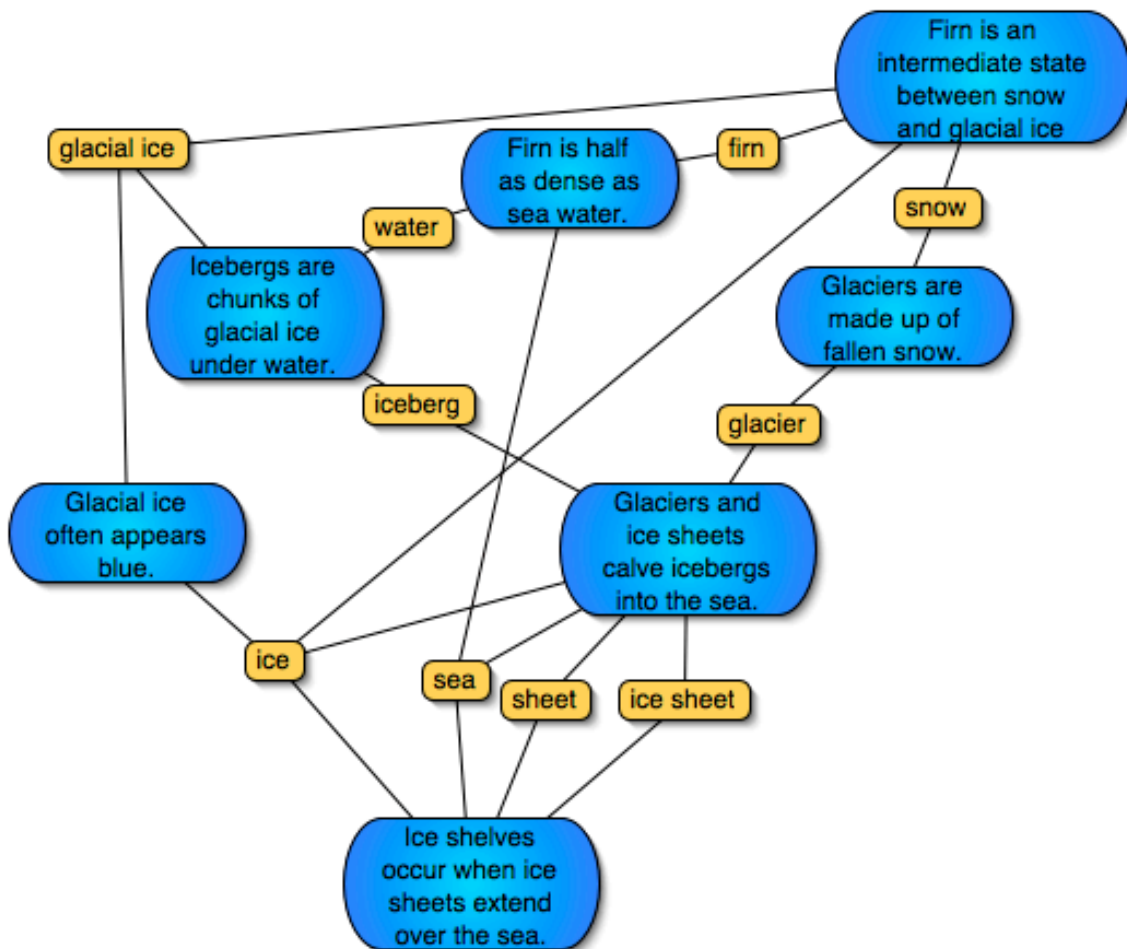


Figure 1: Sample contextual network

CNS represents a significant improvement over LSI as a text search technique. Because document similarity is a function of the number of paths connecting any two nodes in the graph, the search can offer the same kind of expanded recall as LSI, where relevant results come back even without an exact match. Unlike LSI,

however, the approach does not require the same kind of expensive initial indexing step, and does not have problems scaling to very large collections. For collections of exceptional size, the graph may be distributed among many machines, or even moved to the file system, options that are not feasible for a latent semantic search.

A further improvement over LSI is the fact that it is easy to add, remove, or change documents dynamically, without having to re-index the collection. It also becomes easy to incorporate document-document links (such as article citations or hyperlinks), including links generated by domain experts interacting with the system. This feature has no equivalent in vector-space techniques such as LSI, and radically extends the search engine's relevance feedback capabilities. In effect, a CNS engine can learn and improve over time by taking into account user feedback.

Given the advantages outlined here, it is reasonable to wonder what, if any, reasons there are for continuing to work with LSI. The answers lie in the non-textual domain: application of content-based techniques to biological and chemical datasets is so new that no one knows which techniques will work best. The practical equivalence between vector-space and graph-based search techniques observed in text collections may not hold for other types of datasets. For this reason, we will study both families of algorithms.

3. Clustering and Categorization

“Clustering and categorization” describes a family of mathematical techniques. As their names suggest, these techniques are designed to group like items together, ideally in ways that will make sense to human beings. While search algorithms are useful for finding information, categorization algorithms are good at organizing large sets of data into more manageable sets of groups. The goal of clustering is to allow users to quickly get the “gist” of even the largest content collection.

Clustering has been the object of extensive study, and many alternative clustering methods exist, using both graph and vector-space models. Each has its strengths and weaknesses, depending on the nature of the data collection at hand.

NITLE's search and clustering algorithms are tunable, with their behavior controlled by various parameters. Since these settings can have a dramatic effect on search and clustering quality, it is important that we perform empirical tests to find optimum tunings for various data collections. These tests may be automated against a query set with known best results, or performed manually by domain experts.

B. *Experiments on Distributed Processing and Searching across Multiple Databases*

1. Distributed Processing

NITLE's current search algorithms do not impose a theoretical ceiling on collection size, as they can be partitioned among many machines working together in a computing cluster. Collections can also be mirrored across several servers to speed query processing, especially in the case of complex queries consisting of many terms and documents. While the theoretical basis for such distribution is well understood, implementing it correctly is a significant practical challenge. Since a distributed algorithm can allow institutions to make use of existing equipment rather than purchasing specialized hardware, we believe the effort to be worthwhile.

We therefore propose to implement a distributed computing version of the NSE that spreads data structures and processing requirements over multiple machines, using fast Ethernet and low-level communication protocols to link these into a single computing cluster. Once this is achieved, it will be possible to extend the capacity of the NSE by adding commodity hardware, or even running a distributed client in the background on existing computers (such as in an under-utilized computer lab). A manageable initial experiment will turn one 12-workstation computer classroom at the Center for Educational Technology into a computing cluster, and distribute the processing of large datasets over these machines.

2. Peer-to-Peer Search

An important property of the NSE's search architecture is the facility with which instances of the engine can be connected to each other in a network, and their collections searched as if they were a single large collection. This property opens the door to peer-to-peer search and customized search domains, allowing researchers to exercise very fine-grained control by searching multiple document collections across many servers as if they were one unified database. This approach holds great promise for searching distributed data collections, such as locally housed historical materials across the country, or collections where a search needs to cross arbitrary boundaries (such as journal archives). We would draw on the large body of existing theoretical work¹⁰ in designing and implementing the peer-to-peer search. One experiment we plan to do involves searching the University of Virginia's *Valley of the Shadow* archive and Vermont's Addison County historical records of the same historical period.

¹⁰ For example, Powell, J., and Fox, E.A., "Multilingual Federated Searching Across Heterogeneous Collections." *D-Lib Magazine*, September, 1998.
<http://www.dlib.org/dlib/september98/powell/09powell.html>

C. Developing New Technologies and Tools

1. Data Visualization

Data visualization is the technology of representing abstract data using sensory models. Through the use of dynamically generated images, graphs, and charts, very large amounts of data can be displayed and manipulated without overwhelming the user. Interactive elements built into the visual interface allow for more rapid data navigation and discovery than are available with text-based tools. Our visualization efforts will focus on improving the usability and efficiency of the NSE by augmenting textual interfaces with focused and effective data visualization tools. Rather than presenting search results in the NSE only as text, we will experiment with interactive visual formats that can convey much more information quickly to the user. Technologies ranging from script-enabled SVG (Scalable Vector Graphic) files viewable in any browser, to Java applets or Flash plug-ins, can all serve to enhance usability and enable more rapid user feedback.

2. Information Management

We will develop an NSE information management interface which is closely integrated with the semantic search component. This user interface will allow expert users to quickly organize and classify an unstructured data collection. The first time the tool is used, the user will be presented with a visual display showing a list of computer-generated document clusters. From here, the user will be able to rearrange documents, merge clusters, and form new clusters and sub-clusters, creating a personal view of the data collection. Multiple users will be able to create separate views of the collection, so that different users can organize the collection to best fit their own needs. From within the tool, information regarding direct links to other documents (hyperlinks, citations) and inferred semantic links will be available, as will all the NSE search and relevance feedback features. The goal will be to maximize usability and productivity for the domain expert organizing a data collection “from scratch.”

IV. Implementation in Research Domains

A. The Humanities

The Internet has the potential to revolutionize the study of the humanities, giving students and scholars the ability to access copious primary sources from anywhere in the world and rapidly find new material of interest. The lack of powerful search tools, however, has meant that many useful online resources are isolated in “silos” where they are not used to their full potential.

The NSE has the potential to make scholarly resources much more useful to researchers and students alike, both by making them accessible through a simple search interface, and

by allowing users to target searches to constellations of resources they can assemble as a peer-to-peer network.

1. History

UVA *Valley of the Shadow* Articles:

The University of Virginia is continuing to digitize and expand its collection of Civil War–era newspaper articles and official records as part of its *Valley of the Shadow* project¹¹. This material is extensively tagged and available online, but the current site search features are somewhat cumbersome and underused. Ongoing work would involve extending the semantic search pilot project into a comprehensive site search engine, using the NSE as a platform.

Addison County Historical Documents:

The Addison County (Vermont) archive is a digitization project currently being conducted by Middlebury College. The archive will eventually contain texts and images from a variety of local historical sources, spanning the entire history of towns in Addison County. In addition to providing a search engine for these materials, NITLE can use the Civil War–era subset of Addison County documents in conjunction with the UVA archive to create a working demo of a cross-domain, peer-to-peer search engine. This demonstration project could serve as a model for more extensive local history digitization efforts.

2. Literature

Thanks to efforts like the Gutenberg Project¹², as well as countless more specialized efforts by teachers and scholars, there is a wealth of literary text available for online study. Literary text, however, poses some unique challenges in the online context. Most search and publication tools are designed for use with factual, descriptive writing in prose, and not well suited to dense poetic texts or entire novels published online.

NITLE is well positioned to address the special needs of literary scholars by combining the backend of the NITLE Semantic Engine with its existing experimental tools for presenting and annotating literary texts online in a Web application — what we call our literary toolkit.

The Rossetti Archive:

The Rossetti Archive¹³ consists of a wide range of source materials, from manuscripts, to fair copies of poetry, to critical texts. This diversity of text types makes it an interesting demonstration project for the NSE, as well as for NITLE’s

¹¹ “The Valley of the Shadow: Two Communities in the American Civil War.”
<http://www.iath.virginia.edu/vshadow2/choosepart.html>

¹² Gutenberg Project: <http://promo.net/pg/>

¹³ “The Complete Writings and Pictures of Dante Gabriel Rossetti.”
<http://www.iath.virginia.edu/rossetti/index.html>

literary tools. The proposed project would center on building a comprehensive search engine for the Rossetti website, as well as providing a management interface for the archive maintainer through the NSE. Professor Jerry McGann, a 2002 recipient of The Mellon Foundation's Distinguished Achievement Award, and the Rossetti Archive's creator, is eager to collaborate with NITLE on this project.

The Electronic Archive of Early American Fiction:

David Seaman, Director of the Digital Library Federation, formerly of the University of Virginia, has lent his support to the NITLE NSE project, and will facilitate the collaboration of the NITLE development team with UVA's library and Electronic Text Center staff. We hope to use the UVA archive to (a) experiment with comparative searches of tagged data and "raw" data in order to ascertain the differences, if any, in processing performance and search results, and to learn more about the economics of metadata tagging efforts; (b) experiment with user interface design and refinement, with the assistance of UVA's domain specialists.

3. Image Database with Textual Metadata

The University of California, San Diego Image Archive Project:

We have consulted with Linda Barnhart, the head of the Catalog Department and Project Coordinator of the Union Catalog of Art Images (UCAI) Project at the University of California, San Diego (UCSD), and agreed to investigate the possibility of experimenting with the textual metadata that UCAI has in its possession. UCAI will send us a small but typical dataset for NITLE staff to examine, and if the dataset is rich enough for the optimal use of the NSE, then we will work out with UCSD data access details and a protocol for testing larger chunks of UCAI data. We are also very interested in learning about the Content Based Image Retrieval technology with which UCSD is experimenting, and will find out if there are ways to collaborate with UCAI on that aspect of their data.

B. Bioinformatics

When Lee Hood began his career as a young biologist at Caltech in the 1970s, he was so frustrated by the limitations of the computational tools of that time that he divided his lab into two functions — one for pursuing biological research, the other for creating technology to drive the biology forward — essentially creating the field of bioinformatics. The resultant DNA sequencer enabled the success of the Human Genome Project. Today, the 30,000 human genes have been mapped, but scientists must learn how all their "genetic components" work together. The challenge of understanding the interactions among the 30,000 genes and perhaps 300,000 proteins is orders of magnitude more complex than the Human Genome Project. The equivalent of the DNA sequencing instruments are yet to be devised.

NITLE sees the importance of involving liberal arts colleges in this research, and has begun a bioinformatics initiative to assist undergraduate curriculum development in this area. We envision the NITLE Semantic Engine as an integral and significant component of this initiative.

The NSE and Proteomic Analysis:

The *proteome* is the protein complement expressed by the genome of an organism, i.e., all the proteins that are found in a given cell of an organism. The goal of proteomic study is to identify putative proteins and study expression changes, structure, interactions, and functions of proteins on a genome-wide scale. This analysis typically involves the following steps:

- Protein isolation from cells or tissues.
- 2-D electrophoresis to separate proteins according to isoelectric point and molecular weight.
- Detection of protein spots and image analysis.
- I-gel digestion with appropriate protease to generate small peptide fragments.
- Mass spectrometry to generate peptide mass fingerprints (PMF).
- Genomic sequence database searching to match query PMF data with putative ORF (Open Reading Frames).

Many of these steps involve computational analysis, and automation is the key component in increasing the throughput of these repetitious processes. Specifically, the last step above, genomic sequence database searching, is usually carried out using BLAST¹⁴ or FASTA¹⁵, or else done locally using some version of a Hidden Markov Model. These searches are relatively expensive from a computational point of view.

Since the PMF search is essentially a similarity search that tries to find what known proteins are similar to the new ones under investigation, our semantic search approaches are directly applicable. As a proof of concept, we have analyzed a collection of proteins from the Protein Data Bank (PDB)¹⁶ and have achieved good results in clustering the entire dataset. These results were presented at the recent O'Reilly Bioinformatics Conference¹⁷.

Although preliminary experiments show that the NITLE Semantic Engine works considerably faster than the commonly used approaches and produces competitive results, we would like to investigate this further, searching against much larger datasets than we have tried so far.

¹⁴ Basic Local Alignment Search Tool. <http://www.ncbi.nlm.nih.gov/BLAST/>

¹⁵ FASTA. <http://fasta.bioch.virginia.edu/>

¹⁶ Protein Data Bank. <http://www.rcsb.org/pdb/>

¹⁷ "Using Latent Semantic Analysis in Bioinformatics". Conference presentation, O'Reilly Bioinformatics Conference, February 2003. http://conferences.oreillynet.com/cs/bio2003/view/e_sess/3406

C. *Weblogs*

Weblogs first appeared on the Internet in the late 1990s. In their earliest incarnation, weblogs (or “blogs”) often took the form of online diaries, or lists of interesting hyperlinks presented in reverse chronological order. The first weblogs were informal in tone and were often hand-coded in HTML, requiring a level of technical acumen that kept the blogging community small¹⁸.

As Internet tools grew in sophistication, this earliest generation of webloggers began to develop ways to simplify writing on the web. Hosted sites like Blogger¹⁹ and LiveJournal²⁰ made it possible for non-technical newcomers to the Web to start their own online journals, and the phenomenon rapidly gained popularity.

Over the past two years, weblogs have begun to visibly mature and differentiate. There has been a marked growth in specialist weblogs, with professional communities like translators, venture capitalists, and lawyers creating networks of blogs to share knowledge and develop their own contacts. At the other end of the spectrum, young people have been busy creating their own online worlds, where they can discuss their offline lives and interact with their friends in complete freedom.

Much of the current technical innovation on the Internet is taking place within the blogging community, since widely used blogging tools give programmers a way to rapidly try out experimental technologies in the real world. Many of the most sophisticated Semantic Web concepts (such as Web services, or RDF vocabularies²¹) have seen their first and only real deployment in the weblog domain.

Another intriguing aspect of weblogs is the phenomenon’s international reach. Iran and Brazil both have enormous online communities, second only to Anglophone countries in extent. In Iran, blogging has acquired a political aspect, as frustrated young people turn to the Internet to find the kinds of freedoms they are denied in their daily lives²². This spontaneous growth of Web communities around the world opens a fascinating window on other cultures, and creates unparalleled opportunities for learning.

Despite the weblog boom, and the number of tech-savvy people participating in it, there is very little data available about the extent of the phenomenon. Estimates range from a million active weblogs to over five million, but hard numbers are impossible to come by. Because of the frequency of updates, large number of document-document links, and significant size of the dataset, the census will make an ideal test collection for the NSE. Our aim will be to create and maintain a public weblog search engine using the semantic

¹⁸ Blood, Rebecca. “Weblogs: A History and Perspective.”
http://www.rebeccablood.net/essays/weblog_history.html

¹⁹ <http://www.blogger.com>

²⁰ <http://www.livejournal.com>

²¹ W3C Semantic Web. <http://www.w3.org/2001/sw/>

²² Derakhshan, Hossein. “Weblogs, an Iranian Perspective.” Conference presentation, BlogTalk 2003.
<http://hoder.com/weblog/archives/006659.html>

search technologies outlined earlier. Such a project will give us valuable raw data about search engine usage, and serve as an effective test of the scalability and rapid update features we will be building into our software tools. Given the presence of competing weblog search engines, it will also serve as an objective metric of search engine usability and quality.

V. Deliverables

We envision three distinct but related categories of deliverables:

- A. Search and clustering algorithms;
- B. Information management tools and user interface features;
- C. The application of the Semantic Engine in domain areas to provide integrated implementations of the information management tools, interface features, and search techniques.

The development, installation, and testing of these domain-specific implementations will be an iterative process. We expect that interaction with domain experts will lead to changes in the NSE feature set, and possibly exploration of other search algorithms and techniques that have not yet been considered by us. We are also keenly aware that we are describing an effort that spans three years, in the context of rapidly changing technologies. We will be prepared, at all times, to capitalize on opportunities that may unveil themselves to us. This has the necessary consequence of requiring flexible, amendable deliverables and timelines.

Thus, while the categories of the deliverables will remain constant, their actual instantiations may very well change during the course of the three-year period. Our intention is to not only meet the specifications listed below, but to overtake them in technical sophistication, performance, and usability as we continue to refine the system.

A. *Search and Clustering Algorithms*

For textual search and analysis, we will deliver a Contextual Network Search engine that is based on the graph traversal algorithm (described earlier in this proposal), while continuing to work with the Latent Semantic engine, based on singular value decomposition techniques, in bioinformatics. Combined with distributed processing, this set of techniques will give us the ability to scale up to very large collections (e.g., for text, millions of documents).

B. *Information Management Tools and User Interface Features*

1. Auto-Categorization and Archive Management Component

We have so far experimented with auto-categorization and archive management tools as a proof-of-concept project. We will stabilize these prototype features and integrate them into the NSE, creating a uniform Web interface and tying in the

new code with existing search features. The information management component will include auto-clustering, individual user views, and support for integrating user feedback into the search system.

2. Peer-to-Peer Searching

The architecture of the NSE search engine will allow multiple document collections across many servers to be searched as if they were a unified whole. This opens the possibility of many researchers running their own instances of the NSE, and having the ability to connect them together into tailor-made search domains. We will develop a discovery and search protocol to enable the peer-to-peer search feature, along with a user interface for assembling the search results.

3. Visual Data Models

There are many potential data models to explore, including radial displays, hyperbolic projections, and “walkable” graphs. In these models, the size, shape, color, and texture of each document can encode information about the document such as date, author, or reading level. Instead of using text commands to limit their search, users can use visual controls, such as sliders and radio buttons, to directly modify the result set (for example, limiting a search by date). This makes it possible to quickly analyze various branches of the document space, without necessarily running a tedious number of searches. Such visual enhancements hold great potential for increasing the effectiveness and usability of the NSE. We will consult with our domain expert collaborators for the most appropriate visualization models to develop.

4. Graphical User Interface (GUI)

The current version of the NSE uses a Web page as its search interface. We will work to develop an application to serve as a platform-independent GUI for both the search and archive management components of the NSE, for maximum flexibility and ease of use. This interface will include data visualization and navigation controls, support for drag-and-drop, integration with the regular file and print features, and a comfortable look and feel.

5. Stand-Alone Desktop Application

We will develop an implementation of the NSE for individual use, presented as a regularly installed software package requiring no special configuration or expertise on the part of the user. The NSE will index the contents of a special directory, where the user can place text files to be searched. The stand-alone version will offer a browser-based user interface, cross-platform compatibility, and appropriate documentation.

C. *Domain-Specific Tools and Features*

Once we begin working with domain experts, we will have a greater understanding of their specific needs and will design desirable interfaces and tools for them. We envision that the following tools might be useful for several of the specific domains:

1. **The Literary Toolkit**

This will be a variant of the NSE specifically designed for presentation of and interaction with literary texts, retaining all of the NSE's search functionality while offering navigation and annotation tools suitable for lengthy and complex texts.

2. **Visual Navigation and Relevance Feedback**

For existing Web-accessible archives such as the Rossetti site and the Early American Fiction collection, we will include visual navigation and relevance feedback features, designed to maximize usability and access to all materials in the collection.

3. **Blog Census**

We will deliver a Blog Census website, which provides up-to-date statistics on the total number of weblogs found, language and authoring tool distribution, most popular sites by language, and geographical distribution of bloggers. The site will aim to be the authoritative source of information about weblog prevalence, and make raw crawl data available to other researchers studying the blogging phenomenon. As part of the Blog Census, we will keep an archive of crawled sites, taking a "snapshot" of weblogs in our collection every 10 days. Furthermore, we expect the semantic differentiation will lead us to clusters of "knowledgeblog," which are receiving increasing attention from researchers and the general public alike.

4. **Bioinformatics**

In this domain, happily, all the data that we need now are freely available online. We already have the PDB and several other collections resident on our servers. We will, during the grant-funded period, experiment with using the conformational structure as well as the amino acid sequence data to obtain better clustering and search results. Since visual models are crucial to the analysis and interpretation of the data, visualization of the clustering results will be an important component of our investigation.

NITLE has already sponsored one bioinformatics conference for liberal arts colleges, attended by over 60 faculty members. We are in the process of convening meetings with college biologists, computer scientists, and mathematicians to consider ways of introducing bioinformatics in their teaching

and research. We will make NITLE's bioinformatics search and discovery tools available to scientists for experimentation, so that even those scientists with minimal training in bioinformatics will have the ability to conduct important research. They, in turn, will assist us in testing and refining the NSE.

We plan to hire the programming staff as soon as the proposal has been approved, and start implementing A and B above. We will schedule a meeting in early fall 2003 with the University of Virginia administration, library and electronic text center staff, and interested faculty to demonstrate the current version of the NSE, discuss desired feature sets, set up a communications protocol for accessing the *Valley of the Shadow*, Rossetti, and Early American Fiction data, and plan the testing of the system by researchers and/or students. We will also contact UCSD and obtain textual data samples from them to experiment with.

A bioinformatics conference is being planned for spring 2004. We will be able to reach a large number of life science and computer science faculty members from liberal arts colleges at that conference, and begin working with them on their specific needs. Prior to that time, we will have experimented with large datasets from the protein data bank and refined our techniques.

We will deliver the weblog census in early 2004, at the latest, with the features described above. A stand-alone application version of the NSE will be ready for campus deployment by the end of 2004. While we will continue to present our work in progress at national and international conferences and at collaborator sites, we expect 2005 and 2006 to be the period in which a version of the NSE is mature and robust, and we scale our deployment rapidly: in the domains that we work on, the campuses on which the system is installed, and the size and variety of the datasets. We will be guided by our collaborators, users, and our evaluators in our venture. We will present an annual report to The Foundation to summarize the progress made and the funds expended.

The year-by-year "task time-line" in the section on Assessment, below, is provided within the context of necessary flexibility, openness to emerging technologies, and the changing needs of our collaborators and the colleges that NITLE serves.

VI. Project Team

The NITLE NSE team will consist of four key members. The three existing team members are: Clara Yu (Principal Investigator), Maciej Ceglowski (Lead Developer), and John Cuadrado (Consulting Chief Scientist). We will hire an additional half-time programmer who will be responsible for detailed implementation of the system, including stand-alone components, documentation, unit testing, and debugging. NITLE will, as part of its cost-sharing with this project, provide an additional programmer to work with the project team on implementation and testing of the algorithms, user interface, and domain-specific applications. In addition, NITLE will provide administrative and other necessary staff support.

The curricula vitae of the three existing team members are included as Appendix A.

VII. Assessment and Timeline

The proposed technologies will be tested in the domain areas described by collaborators from academia and the research community and through intense interaction with users (e.g. faculty, and students doing research under faculty supervision). We have always benefited from user feedback, presented our work at national conferences, and reacted quickly to rapidly changing technology. These “market” reactions to our products will serve as stringent and real-time assessments of our work.

In addition, we are very fortunate to have three nationally known scholar-researchers as our evaluators.

Michael Keeble Buckland is professor of Information Management and Systems at the University of California, Berkeley. He has a strong background in library science and is one of the international authorities on library systems, organization of knowledge, and digital resources management. He brings to the NSE projects his expertise in the history and theory of documentation, extensive knowledge of the state of the art in bibliography, cataloging, metadata, and retrieval, and a curiosity and open-mindedness that we admire and emulate.

Robert A. McCaughey is professor of History at Barnard College, and serves on the Columbia University graduate faculty. He has served as academic dean at Barnard, and now directs Barnard’s Electronic Archive and Teaching Lab. He is also a member of the NITLE Board of Directors. Bob brings with him intimate knowledge of NITLE’s mission, an understanding of the liberal arts college environment, a deep appreciation of the issues around digital repositories, and an eagerness to work with American history research materials, such as the *Valley of the Shadow* archives.

John M. Unsworth has been the director of the University of Virginia’s Institute for Advanced Technology in the Humanities (IATH). He has recently been appointed dean and professor of the Graduate School of Library and Information Science, as well as professor of English, at the University of Illinois (Urbana-Champaign). John is generally regarded as “Mr. Digital Humanities,” and has done world-class work in pushing humanities scholarship into the digital realm. He has been most encouraging to NITLE in its development of the semantic engine, visualization and archivist’s tools, and has been instrumental in forging collaborative relationships between NITLE and the University of Virginia.

Curricula vitae for the three evaluators are included as Appendix B.

The evaluators will meet at least once a year with the project team to see demonstrations of the NSE in its current state of development and deployment, and to discuss its progress. In addition, the project team will present progress reports, answer questions, and seek advice from the evaluators periodically. In the second half of the third year,

each evaluator will produce a report summarizing his findings. These reports will be included in the final NSE report to The Foundation.

Within the context of the flexibility and openness to new technologies described in the section on deliverables, the project team will work with the following timeline, against which the evaluators will gauge the project's progress:

2003-2004 TASKS	2004-2005 TASKS	2005-2006 TASKS
Implement NSE with search and clustering algorithms, information management tools, and a graphical user interface.	Refine search and clustering algorithms, information management tools, and the graphical user interface.	Continue to refine system. Deliver stand-alone application. Implement distributed computing.
Consult with collaborators and plan for domain-specific implementation with specialized features for literary analysis, visual navigation, relevance feedback, blog searching, and bioinformatics.	Continue consultation with collaborators. Deliver Blog Census version 1.0.	Increase collaborator base and domains. Continue to develop Blog Census, possibly implement a catalog of knowledgeblogs.
Experiment with <i>Valley</i> , Rossetti and Early American Fiction data.	Deliver the NSE to collaborators for testing and research: <i>Valley</i> , Rossetti, and Early American Fiction projects.	Implement peer-to-peer searches across databases.
Experiment with UCSD textual metadata (if appropriate).	Hold bioinformatics conference. Develop and deploy bioinformatics tools.	Continue to develop and deploy bioinformatics tools.
Implement the literary analysis tool.	Develop stand-alone application version of NSE.	Increase "installed base" on college campuses.
Present at conferences and collaborator sites.	Present at conferences, collaborator sites, and on NITLE campuses.	Present at conferences, etc. Possibly hold conference to showcase collaborator projects. Write and submit final report.

VIII. Proposed Budget

A. Project Staff

For the development and refinement of algorithms, programming, documentation, testing and deploying the system, collaborating with researchers in various domains, presenting and demonstrating the NSE system at conferences, and working with assessment experts and evaluators, we propose the following salaries and benefits for project team and staff members, to be covered by the budget of this grant.

1. Three years' salary and benefits for the Consulting Chief Scientist (2003-6), who is responsible for integrating mathematical algorithms into NSE systems with the aim of continually refining the system and improving its performance.
2. Three years' salary and benefits for the Lead Developer (2003-6), who will be in charge of systems development, integration of new algorithms into the system, peer-to-peer computing, and distributed processing.

3. Three years' salary and benefits for a .5 FTE Staff Programmer, who will be responsible for detailed implementation of the system, including stand-alone components, documentation, unit testing, and debugging.

The salary and benefits of the project's Principal Investigator (.25 FTE) and an additional 1 FTE Programmer will be funded through a NITLÉ cost-share.

B. *Demonstration and Dissemination*

1. Expenses for presentations and demonstrations of the NITLÉ Search Engine to collaborating institutions.
2. Travel to national and special interest conferences, such as the O'Reilly Emerging Technologies Conference, the Coalition for Networked Information conferences, and the Bioinformatics Technology Conference.

C. *Hardware and Storage*

1. Development of the NITLÉ Search Engine will require a server that meets the following specifications:

ABERDEEN STIRLING S47 - 4U RACK

Part#	Description - Quantity
MB6862	SuperServer 7042-P8RB Xeon SCSI Server (blk) - 1
MI5585	Rails for SC742 Rack MB6828 (black) - 1
IC7809	Single Intel Xeon 2.8 GHz 400FSB - 2
MYC289	1GB Corsair PC2100 Reg ECC DDR SDRAM - 8
IS121	LSI MegaRAID U320 1-Ch. RAID Controller - 1
HDG7310	Seagate Cheetah10K 73GB 4.7ms Ultra320 SCA 8MB - 7
TP423	Quantum SDLT 160/320GB Tape Drive - 1
MI927	CABLE LVD 68M TO 68M 3FT W/ TERMINATOR - 1
HCR151	SONY CRX220E1 52X24X52X Black CD-ReWriter - 1

In the second year, we plan to add storage as follows:

ABERDEEN BLACKWATCH S38i - 3U RACK

Part#	Description - Quantity
KITS38i	**Aberdeen BlackWatch S38i Kit** - 1
IC453	Boxed Pentium 4 3.06 GHz/512K/533FSB processor - 1
MYDD206	512MB ECC REG DDR266 SDRAM 184-pin DIMM - 2
HDG25001	Maxtor 250GB 9ms 5400rpm ATA/133 2MB - 8
BPS38M	Maxtor IDE Backplane (CS130) - 1

2. Two high end workstations for project staff, configured as follows:

ABERDEEN INVERNESS W48 WORKSTATION

Part#	Description - Quantity
MB6881	SuperWorkstation 7043A-8RB 4U Xeon SCSI (blk) - 1
IC7820	Single Intel Xeon 2.8 GHz 533FSB - 2
MYC289	1GB Corsair PC2100 Reg ECC DDR SDRAM - 6
IS948	Adaptec 2010S 64-bit LP PCI Ultra320 ZCR - 1
HDG14002	Seagate Cheetah10K 146GB 4.7ms Ultra320 SCA 8MB - 3
DAG951	PNY NVIDIA Quadro4 980XGL AGP 8X 128MB DDR - 1
HCR132	Plextor PlexWriter 12X/10X/32X SCSI ReWriter - 1

3. A half-rack of off-site data storage and backup will be obtained at a cost of \$400/month.

The above equipment will be supplemented by four computers (two desktops, two laptops, and accessories) to be provided by NITLÉ as part of its cost-share. These machines will be used by project staff (PI, programmer) and during staff travel.

D. Assessment

We have requested funds to cover annual visits to NITLÉ by each of the three evaluators as described in the section on Assessment, and to provide them with honoraria upon completion of their final reports.

E. NITLÉ Cost-Share

In addition to the personnel and equipment costs mentioned above, NITLÉ will provide office space and some equipment for the project.

IX. Conclusion

This is an exciting time for us to be involved in the exploration, analysis, and production of knowledge. On the one hand, we have the information explosion and the ubiquity of communications access; on the other hand, we have the real problem of managing the process and product of this revolution.

NITLÉ, as an experimental networked virtual organization in higher education, has had the good fortune to be supported by The Foundation in its transformation of theoretical

research results (such as linear algebra and graph theory) into applications for information management. In the course of our work, we receive constant reminders, through feedback received from our collaborators, that a small team of people can have a great effect on the work of many. This is probably the fundamental reason why the core NSE staff works with a spirit of constantly heightened elation.

We have also found that our enthusiasm is contagious. At conferences and presentations, we are inundated with queries about the availability of the NSE and the possibilities for collaboration. The proposed development and testing activities will enable us to answer these requests, and answer them well.

We are keenly aware of the rapid speed with which new technologies come to market. We realize that everything that is written in this proposal represents only our current vision, understanding, and community of collaborators. Things will almost certainly change; we will almost certainly work with more institutions and individual researchers than we have identified here; we might even continue discovering and applying better algorithms for better search, information management, and visualization results.

We are grateful to The Foundation for its support of this quest, and for its vision and leadership.

X. Appendices

A. *Project Team Curricula Vitae*

1. Maciej Ceglowski

Lead Developer, National Institute for Technology and Liberal Education.

Academic Degrees

Middlebury College, A.B., Russian and Studio Art (double), 1997. *Summa cum laude*, Phi Beta Kappa.

Fellowships

Elizabeth Greenshields Foundation (Oil painting). 1998 and 1999.

Thomas J. Watson traveling fellowship. "Oil Painting in Northern and Southern Light." 1997-98.

Professional Interests

Natural language processing: work to date has focused on topic boundary identification, part-of-speech tagging, stemming, statistical language identification, and noun phrase extraction.

Information retrieval. Design and implementation of search engines for natural language and scientific data collections, including keyword, latent semantic analysis, and spreading activation algorithms.

Web application design. Development of scalable, fast Web applications on the LAMP (Linux + Apache + MySQL + Perl/PHP) platform.

World Wide Web: Exploration of linking patterns in Web communities, particularly across language barriers.

Publications

"Building a Vector-Space Search Engine in Perl", perl.com website. (<http://www.perl.com/lpt/a/2003/02/19/engine.html>) February 2003.

"NITLE Experiments with New Search Technologies," nitle.org website, (http://www.nitle.org/article_new_search.php). August 2002.

“Patterns in Unstructured Data: Discovery, Aggregation and Visualization.” A presentation to The Andrew W. Mellon Foundation, published online. (http://javelina.cet.middlebury.edu/lisa/out/cover_page.htm), June 2002.

Published Software

All of the following are Perl modules available on the Comprehensive Perl Archive Network (CPAN).

WWW::Blog::Identify – Heuristics for identifying weblog authoring tools.

Lingua::EN::Tagger (with Aaron Coburn) – Part-of-speech tagger for English text.

Search::VectorSpace – Simple vector model search engine.

Search::ContextGraph – Perl implementation of contextual network search.

Conference Presentations

“Managing Unstructured Data with Latent Semantic Indexing.” CNI, April 2003.

“Peer-to-peer semantic search engines: building a Memex.” O’Reilly Emerging Technology Conference, April 2003.

“Using Latent Semantic Analysis in Bioinformatics.” O’Reilly Bioinformatics Conference, February 2003.

“Intelligent Searching of Media Databases.” NERCOMP, December 2001.

2. **John L. Cuadrado**

Project Consultant, National Institute for Technology and Liberal Education
550 Hinesburg Road, Suite 302
South Burlington, VT 05403

Primary Areas of Interest

Image processing, signal processing data flow systems, computer vision, software-hardware codesign, neural networks, modeling and simulation, distributed systems, fuzzy logic, non-monotonic reasoning, client-server database systems.

Education

BS in Physics, 1969. Fairleigh Dickinson University, Teaneck, NJ.
MS in Physics, 1972. University of Illinois, Champaign, IL.
Ph.D. in Mathematics, 1977. University of Illinois, Champaign, IL.

Professional Experience

January 90 to present: Independent consultant

October 81 to December 89: President of Octy, Inc.

During the past 14 years I have worked as an independent consultant. I am currently working on various implementations of distributed versions of a new generation of Markov Chain Monte Carlo methods that are of use in image processing and many other applications. An important component of this development effort has been the modeling of these distributed systems. Both analytical models as well as simulation techniques were developed. I have extensive background in stochastic modeling going back to my work in the DOD's VHSIC program in the early 1980s. I have also developed a large expert system that applies fuzzy logic to the field of oriental medicine. The system helps practitioners develop a diagnostic and provides recommendations on acupuncture points and Chinese herbs. These current projects have all been developed using MS Windows, MS Visual C++, and Visual Basic with MS Access serving as the database engine.

From 1984 to 1990 I was involved in the design and implementation of the Big Floyd and OBR III expert systems for the FBI. These systems are designed to aid agents in their missions against organized crime and terrorism. Both systems are extremely large and use state of the art artificial intelligence techniques, some developed specifically for these two systems. Both systems run against large intelligence databases maintained by the FBI. The rules used by these two expert systems have been obtained directly from FBI special agents working actual cases. The rule acquisition process has been an ongoing activity for the past six years and has produced several thousand Prolog-based rules that the systems use in analyzing criminal enterprises. My role in these systems was as chief scientist and principal designer. Modeling of system capacity and response times was an

integral part of the development of these expert systems. These expert systems were developed using Sun workstations and were written in a combination of Prolog and C using Sybase as the database backend.

Served as principal investigator in a Department of Defense sponsored study of applications of artificial intelligence to design automation.

Served as principal investigator on the design of a high-performance inference engine for the efficient execution of logic programming constructs and rule-based systems.

Designed and developed a data flow methodology, referred to as the Directed Graph Methodology (DGM), which was used in a number of programs including VHSIC, E-3A and AOSP. Developed a natural language front-end for the DGM design tools. Participated as principal consultant on Research Triangle Institute's ADAS (Architecture Design and Assessment System) design and implementation.

Designed the operating system to be used in the ATF VHSIC brassboard. Designed an extended ISA for the 1750A to provide efficient support for the memory management and multi-tasking facilities required in an Ada implementation. Designed a fully distributed OS support for DGM used in the Westinghouse AOSP study program. Developed process/processor binding algorithms for distributed systems in general.

February 81 to October 81: Technical Director of Signal Processing, SofTech Inc., Falls Church, VA.

Provided technical direction and customer interface in the ASW Common Operating System (ACOS) program. Served as manager and technical lead in the preparation of all signal processing proposals and white papers. Was responsible for the original version of the ECOS data flow methodology which has become a Navy Standard and will be used as the programming vehicle in the Navy's Enhanced Modular Signal Processor (EMSP).

June 80 to February 81: Senior Systems Analyst, Lexico Enterprises, Washington, DC.

Was a principal member of the design team for the Boeing Pneumatic Atlas Compiler. This compiler was designed to be used as the vehicle for writing the test procedures for the Avionics subsystems in the Boeing 757, 767 and the BMAC retrofit of the B52. Primary responsibilities included: the design and direction of the implementation of the code generation and semantic analysis phases of the compiler, the specification and design of various preprocessors used in other phases of the compiler.

September 78 to June 80: John Wesley Young Research Instructor, Department of Mathematics, Dartmouth College, Hanover, NH.

August 77 to August 78: Visiting Assistant Professor, Department of Mathematics, Wright State University, Dayton, OH.

Publications

Cuadrado, J. L., "Parallel Query System", BYTE, Jul. 1995.

Cuadrado, J. L., "Mining Statistics", BYTE, Feb. 1995.

Cuadrado, J. L., "Teach Formal Methods", BYTE, Dec. 1994.

Cuadrado, J. L. and Pimentel, S. G., "A Prolog Implementation of the Stable Model TMS", Proceedings of RMCAI-90.

Cuadrado, J. L. and Pimentel, S. G., "The Event Calculus and Consistency Maintenance", Proceedings of the Third International Conference on Industrial Engineering Applications of Artificial Intelligence and Expert Systems, 1990.

Cuadrado, J. L. and Pimentel, S. G., "A Truth Maintenance System Based on Stable Models", Proceedings of the North American Conference on Logic Programming, The MIT Press, 1989.

Cuadrado, J. L. and Pimentel, S. G., "A Horn Clause Theory of Inheritance and Temporal Reasoning", Lecture Notes in Artificial Intelligence, Vol 390, Springer Verlag, 1989.

Cuadrado, J. L. and Linsenmayer, G. R., "GTS: An Expert System for Graph Transformations", Proceedings of the 2nd International Conference on Industrial and Engineering Applications of AI and Expert Systems '89.

Cuadrado, J. L. and Yu, C. "Vision Systems, Programming", pp 1904-1915 in "International Encyclopedia of Robotics Applications and Automation", John Wiley & Sons, 1988.

Cuadrado, J. L., Cohen, B., and Kendall, K. "Intelligent System for Analog Design," "Proceedings, IEEE CompCon '86", pp 118-120.

Cuadrado, J. L. and Cooley, E. S., "ISSD an Intelligent System for DSP design," "Proceedings, IEEE CompCon '86", pp 121-123.

Cuadrado, J. L. "Design Automation Software Tools: The State of the Art", "Proceedings, Aerospace Applications of AI '86".

Cuadrado, J. L. and Cuadrado, C. Y. "Handling Conflicts in Data", Byte, November, 1986.

Cuadrado, J. L. and Cuadrado, C. Y. "AI in Computer Vision," Byte, January, 1986.

Cuadrado, J. L. and Cuadrado, C. Y. "Logic Programming Goes To Work," Byte, August, 1985.

Cuadrado, J. L. Artificial Intelligence Techniques for Integrated Design Automation, prepared for the VHSIC program office, Wright Patterson AFB under contract to UES, Dayton OH, January, 1985.

Cuadrado, J. L. "Metalevel Reasoning in the Design of Signal Processing Systems," Proceedings, IEEE CompCon '85.

Cuadrado, J. L., with Frank, G. A. and Smith, C. U. "An Architecture Design and Assessment System for Software/Hardware Codesign," in "Proceedings, IEEE Design Automation Conference, June, 1985.

Cuadrado, J. L. The Implementation of a Prolog Interpreter in Ada, prepared for Westinghouse Defense Electronics Systems Center, July, 1984.

Cuadrado, J. L. Natural Language Parsing Techniques, prepared for Westinghouse Oceanic Division, June, 1984.

Cuadrado, J. L. Text/Discourse Comprehension, Generation, Summarization, prepared for Westinghouse Oceanic Division, August, 1984.

Cuadrado, J. L. with Hinkey, M. and Glaser, B. "An Integrated Tool Set for the Development of Advanced Signal Processing Software," Proceedings, IEEE 1984 National Aerospace and Electronics Conference, Dayton, Ohio.

Cuadrado, J. L. A Review of Expert System Technology and Its Application to the VHSIC Integrated Design Automation System, prepared for the VHSIC/IDAS Steering Committee, Sept., '83.

Cuadrado, J. L., Hinkey, M. Gaertner M. "Microcode Support for Operating System Functions: Issues and Examples," in Proceedings, the Entity-Relation Model Conference, North Holland Publishing Co., '83.

Cuadrado, J. L., Linsenmayer G. R. "Efficient High Speed Implementation of Directed Graph Signal Processing in a Distributed Processing System," in Proceedings of the IEEE CompCon '83.

Cuadrado, J. L., Honey, W. F., Wenk A. F. "System Programming Aids (SPA) Methodology," in Proceedings of the IEEE 1982 National Aerospace and Electronics Conference, Dayton, Ohio.

Cuadrado, J. L. VHSIC Local Operating System Specification, 9RA8742, Westinghouse Electric Corporation, Baltimore, Maryland.

Cuadrado, J. L. VHSIC Kernel Operating System Specification, 9RA8754, Westinghouse Electric Corporation, Baltimore, Maryland.

Cuadrado, J. L. A New AOSP Nodal Operating System for the Efficient Implementation of a Data Flow Methodology, Westinghouse Electric Corporation, Baltimore, Maryland.

Cuadrado, J. L. Directed Graph Methodology, Westinghouse Electric Corporation, Baltimore, Maryland.

Cuadrado, J. L. "Iterated Integrals and Formal Power Series Connections," in Notices of the American Mathematical Society, December, 1977.

Cuadrado, J. L. "Formal Power Series Connections and Rational Loop Space Homology," in Notices of the American Mathematical Society, December, 1976.

3. Clara Yu

Cornelius V. Starr Professor of Linguistics and Languages
Director, Center for Educational Technology
Middlebury College
Middlebury, VT 05753

Director, National Institute for Technology and Liberal Education
550 Hinesburg Road, Suite 302
South Burlington, VT 05403

Primary Areas of Academic Interest

- Instructional technology: design and implementation, collaboration.
- Second language acquisition.
- Latent semantic analysis for data-mining.
- Artificial intelligence: expert system development, natural language understanding and processing, natural language parsing and generation.
- Comparative literature: theories of comparative literature, comparative study of Chinese and Western literature.
- Chinese literature: classical and modern drama, fiction, poetry.
- Chinese culture: society and civilization.
- International studies.
- Creative writing.

Special Projects

In 1994, headed the Task Force for Information Technology at Middlebury College, conducted an audit and formulated a strategic plan for the College's information technology future.

As Vice President for Languages (1993-1996), restructured the operation comprised of the eight Middlebury Language Schools and five Schools Abroad (Florence, Madrid, Mainz, Moscow, Paris), eliminating managerial layers, consolidating staff, and instituting team decision making.

In 1995, designed Middlebury's three-year International Major, a rigorous B.A. program that combined an international studies core curriculum, regional concentration, language competence, study abroad, and senior capstone experience. This innovative program attracted the attention of such national publications as the *Boston Globe*, *Wall Street Journal*, and *Washington Post*. It also served as the prototype of the College's four-year International Studies Major.

Founded the Center for Educational Technology (CET) at Middlebury College, and oversaw the successful completion of its first major project, Project 2001— a five-year initiative to increase the efficiency and effectiveness of language instruction through the use of technology.

Designed Project 2001's six integrated programs for networking higher educational

institutions in building a human infrastructure for curriculum design and delivery, based on a learner-centered, technology-enhanced model. This project linked research universities, graduate programs, and 62 liberal arts colleges in an effort to create an unbroken chain between cutting-edge research, producers of next-generation faculty, and existing faculty and technical staff in the common mission of improving education and reducing costs.

In support of Project 2001, obtained the largest grant ever awarded to a liberal arts college by The Andrew W. Mellon Foundation, \$4.7 million, for a total of \$7 million in developing language and technology programs.

Managed CET's transition from a focus on language learning to a broader curricular mandate to facilitate the effective use of technology for 37 colleges in the Mid-Atlantic and New England Region. Obtained a grant of \$2.1 million for the CET from The Andrew W. Mellon Foundation in support of this initiative as a component of the "Centers Strategy."

While continuing to direct CET, serve as founding director of the National Institute for Technology and Liberal Education (NITLE), which coordinates the activities of three regional technology centers (the Mid-Atlantic and New England region, the South, and the Midwest) and serves as a catalyst for innovation and collaboration for 81 national liberal arts colleges as they seek to make effective use of technology to enhance teaching, learning, scholarship, and information management. This initiative is funded by The Andrew W. Mellon Foundation through a grant of \$4 million administered by Middlebury College.

Education

- B.A.: English, National Taiwan University, Taipei, Taiwan, 1971.
- M.A.: Comparative Literature, University of Illinois, Urbana, Illinois, 1973.
- Ph.D.: Comparative Literature, University of Illinois, Urbana, Illinois, 1978.

Professional Experience

Teaching

Cornelius V. Starr Professor of Linguistics and Languages, 1996 - present.

Academic Administration

Vice President for Languages and Director of the Language Schools, Middlebury College, 1993 - 1996.

Director, Mellon Initiative in Teaching Languages with Technology, 1994 - 1997.

Director, Project 2001, a 62-college consortium for technology-enhanced foreign language curriculum development and delivery, 1997 - 2001.

Director, Center for Educational Technology, 1997 - present.

Director, National Institute for Technology and Liberal Education, 2001 - present.

Publishing

Editor, Chung-Wai Literary Magazine, Taipei, Taiwan, 1971-72.

Associate Editor, Modern Literature, Taipei, Taiwan, 1971-72.

Consulting

Artificial Intelligence Analyst, Octy, Inc., Fairfax Station, VA, 1983-1987.

Primary areas of responsibility: design and implementation of expert systems and natural language processing systems, design of man-machine interface, text generation and summarization.

Served on a team of consultants for the American University in Paris, focusing on the integration of technology, November 2001.

Served as a consultant to the Budapest University of Technology and Economics, Hungary, advising the University on strategic planning in an environment of globalization of higher education, internationalization of the curriculum, facilitation of life-long learning, and the establishment of an institutional research office and an institutional advancement office, April 2002.

Publications and Presentations

Literary Criticism

The “Yueh Fei Theme in Chinese Fiction,” presented at the New England Regional Conference of the Association for Asian Studies, Storrs, Connecticut, 1979.

“Cross-cultural Currents in the Theatre,” in William Tay (ed.), China and the West: Comparative Literature Studies (Hong Kong: The Chinese University Press, 1980), pp. 217-37.

“Modern Chinese Drama after the Cultural Revolution: Realism vs. Idealism,” presented at the Mid-Atlantic Regional Conference of the Association of Asian Studies, College Park, Maryland, 1981.

“The Enclosure Motif in Modern Chinese Fiction,” presented at the Pacific Regional Conference of the Association for Asian Studies, Honolulu, Hawaii, 1982.

“On ‘For John, Who Begg Me Not to Enquire Further’ by Anne Sexton,” in Robert Pack and Jay Parini (eds.), Touchstones (Middlebury College Press, 1996), pp. 310-4.

Chinese Literature

“Parting” (a translation, with Vivian Hsu), in Vivian Hsu (ed.), Born of the Same Roots (Indiana University Press, 1981), pp. 81-93.

“Portrait by a Lady: The Fictional World of Ling Shuhua,” in Angela Jung Pallandri (ed.), Women Writers of 20th-Century China (University of Oregon Press, 1982), pp. 41-62.

Six entries on dramatists of the Ming and Ch’ing eras: “Hsi Yung-jen,” “Kuei Fu,” “Shen Tzu-cheng,” “T’ang Hsien-tsu,” “Wu Ping,” “Yu T’ung,” in William H. Nienhauser (ed.), The Indiana Companion to Traditional Chinese Literature (Indiana University Press, 1986), pp. 420-1, 511-2, 679-80, 751-2, 900-1, 939-40.

Chinese Culture

Twenty-six sections in Patricia Ebrey (ed.), A Sourcebook of Chinese Society and Civilization (New York: The Free Press, 1982), ranging from a 4th-century biography to 20th-century social records, pp. 47-52, 66-67, 79-83, 100-20, 130-2, 136-40, 145-54, 161-6, 171-5, 189-99, 209-10, 219-23, 135-42, 149-53, 259-60, 269-88.

Language Pedagogy

“Computer-Aided Instruction for Chinese,” presented at the Annual Meeting of the Chinese Curriculum Consortium, Columbus, Ohio, March, 1989.

“Computerized Testing and Tutoring,” presented at the Annual Meeting of the Chinese Language Teachers Association, Boston, MA., Nov., 1989.

Chinese Pronunciation Tutor, HyperGlott Software Company, 1990.

Chinese Survival Manual and Language Laboratory, HyperGlott Software Company, 1990.

Chinese Writing Tutor, HyperGlott Software Company, 1991.

Practical Intermediate Chinese, Field Test Version 3.0, 1991. 278 pp.

“Designing Pedagogically Valid CAI Material for Chinese,” invited lecture, Indiana University, July 13, 1991.

“Pedagogy-Driven Computer-Assisted Instruction for Chinese,” presented at the Annual Meeting of the American Council on the Teaching of Foreign Languages, Washington, DC, November 1991.

Panel chair for “Articulation in Less Commonly Taught Languages,” at the Northeast Conference, New York, NY, April 1994.

Panel chair for “Implications of the National Standards movement for the Less Commonly Taught Languages,” at the Northeast Conference, New York, NY, April 1995.

A Guide for Basic Chinese Language Programs (with Y. O. Biq, G. Henrichson, C. C. Kubler, G. Walker, A. R. Walton, M. Wong, W. Wu), to be published by Ohio State University Foreign Language Publications, 128 pp., 1997.

“Real and Virtual Immersion for Language Learning,” presented at the conference on “Implications of Different Learning Styles for Teaching,” University of Cincinnati, April 11-12, 1997.

“Colorless Green Ideas...,” Inaugural Lecture, C. V. Starr Professorship, Middlebury College, April 1998.

Natural Language Processing, Artificial Intelligence

Natural Language Parsing Techniques, a report prepared for Westinghouse Electric Corporation, 1984. 125 pp.

Text/Discourse Comprehension, Generation, Summarization, a report prepared for Westinghouse Electric Corporation, 1984. 121 pp.

The Use of Frames in Expert Systems, a report prepared for the Institute for Defense Analyses, February 1986. 26 pp.

“Vision Systems, Programming” (with J. L. Cuadrado), in International Encyclopedia of Robotics Applications and Automation (John Wiley & Sons, 1988), pp. 1904-5.

Directed research, sponsored by the National Institute for Technology and Liberal Education, on the discovery, aggregation, and visualization of patterns in unstructured data, and the integration of search technologies based on this research in tools to enhance learning. Organized and participated in numerous demonstrations of this research, at venues such as The Mellon Foundation and the University of Virginia.

Information Technology, Language Pedagogy, and Technology

“High Tech in Higher Education,” Presidential Inauguration Lecture, Middlebury College, November, 1990.

“Technology at the Cutting Edge: Implications for Second Language Learning,” Language for a Multicultural World in Transition (Lincolnwood, IL: National Textbook Company, 1992), pp. 165-195.

Information Technology for the Liberal Arts College of the 21st Century, Report of the Task Force on Information Technology at Middlebury College. February, 1994. 28 pp.

“Rethinking Language Pedagogy with Technology in Mind,” keynote address, Mellon Workshop on Using Technology in Language Instruction, June, 1994.

“An Acquisition Model of Language Pedagogy: Possibilities with Technology,” keynote address, Mellon Workshop on Language and Technology, July, 1995.

Panel Chair, “Technology and Learning: From the Caves of Prehistory to the Frontiers of Cyberspace,” the 82nd Annual Meeting of the Association of American Colleges and Universities, Washington, D. C. January, 1996.

“Global Learning Environments,” presentation made at the 82nd Annual Meeting of the Association of American Colleges and Universities, Washington, D. C. January, 1996.

“Project 2001 and Beyond,” keynote address, Conference on “Technology and Language Instruction for the 21st Century,” Middlebury, June, 1997.

“Language Teaching with Technology: Past, Present, and Future,” Speech at Opening Plenary, Language Conference at University of Puget Sound, Tacoma, WA, Oct. 30, 1998.

“Connections, Collaboration, and Consortia: Project 2001,” Speech at Opening Plenary, Language Conference at Willamette University, Salem, OR, Feb. 8, 1999.

“Language Instruction with Technology,” Speech at Opening Plenary, Language Conference at Whitman College, Walla Walla, WA, Feb. 12, 1999.

“Efficiency and Effectiveness in Language Instruction,” Speech at Opening Plenary, Lewis and Clark University, Portland, OR, Feb. 15, 1999.

“Language Instruction in Liberal Arts Colleges in the 21st Century,” Keynote speech at the Colby-Bates-Bowdoin Language Conference, Colby College, Watertown, ME, April 17, 1999.

“A Brave New Century,” Keynote speech at the New England Regional Language Conference organized by Connecticut-Trinity-Wesleyan, Middletown, CT, May 24, 1999.

“Technology and Liberal Arts Education,” and essay for Middlebury’s bicentennial Founder’s Day Symposium, “Higher Education, the Market, and Media,” Middlebury College, Middlebury, Vermont, November, 2000.

Keynote Address at the Associated Colleges of the South Conference on Information Fluency, Southwestern College, Georgetown, Texas, February, 2001.

Keynote address at Emory University workshop on language pedagogy and technology, Atlanta Georgia, February 2001.

Participated in a satellite broadcast roundtable discussion, “Technology: Tool or Method?” at Duke University, sponsored by the Charles E. Culpeper Foundation, the Rockefeller Brothers Fund, Washington and Lee University, and Duke University, April, 2001.

Keynote at Associated Colleges of the Midwest technology conference on “Technology and the Liberal Arts,” Grinnell College, Grinnell, Iowa, June, 2001.

Plenary address at the inaugural conference of the National Institute for Technology and Liberal Education (NITLE) in Atlanta, GA, January 11, 2002.

International Studies

“Study Abroad Issues for Liberal Arts Colleges,” feature presentation of the NESCAC Special Meeting on Study Abroad, April 1995.

Arab Culture and Civilization, August 2002. Directed the development of this comprehensive collection of educational resources on the Arab world, produced by the National Institute for Technology and Liberal Education (NITLE) in response to the scarcity of curricular materials, and in light of increased interest in this area among students and faculty since September 11, 2001.

Creative Writing

“Dedicated to the Master of Silent Garden,” Spoon River Quarterly, XVI, Nos. 1 & 2, 1991, pp. 30-31.

“Wang Zhaojun,” Spoon River Quarterly, XVI, Nos. 1 & 2, 1991, p. 32.

To the Interior (Selected Poems), Taipei: Bookman Books, Ltd., 1992, 81 pp.

“Five Poems by Clara Yu,” The Chinese Pen (Taipei: International P. E. N.), Vol. 21, No. 1, 1992, pp. 183-90.

“June Elegy,” in Robert Pack (ed.), The Bread Loaf Anthology of American Identities (Middlebury College Press), 1994, p. 358.

“Virgin River Bride,” in Robert Pack (ed.), The Bread Loaf Anthology of American Identities (Middlebury College Press), 1994, p. 359-60.

“Skull,” in Robert Pack (ed.), The Bread Loaf Anthology of American Identities (Middlebury College Press), 1994, p. 361.

“Swallows of Xi'an,” in Robert Pack (ed.), The Bread Loaf Anthology of American Identities (Middlebury College Press), 1994, p. 362.

“A Destitute Time,” in Robert Pack and Jay Parini (eds.), Introspections: American Poets on One of Their Own Poems (Middlebury College Press), 1997, pp. 312-318.

B. *Curricula Vitae of Evaluators*

1. Michael Keeble Buckland

Professor, School of Information Management & Systems
Co-Director, Electronic Cultural Atlas Initiative
University of California, Berkeley

Education

Oxford University, England, B.A., Modern History, 1963.
Sheffield University, England, Postgraduate Diploma in Librarianship, 1965; Ph.D.,
Social Sciences (Economic Analysis and Librarianship), 1972.

Dissertation: *Library Stock Control*.

Special Interests

Design and management of academic library services. Bibliography, filtering, retrieval.
History and theory of documentation and information management. Information and
cultural heritage.

Teaching

Academic libraries. Library management. Organization of knowledge: Bibliography,
cataloging, metadata, retrieval. Information systems. Information and cultural heritages.

Recent instructional improvement activity

Instructional unit on diversity issues in information retrieval (With C. Nolan). (COT
Minigrant, 1992).
IMS 101 *Information systems*. Introduction to ideas about information and information
systems for undergraduates. Introduced 1993.
American Cultures Summer faculty fellowship, 1994.
IMS 142 *Access to American cultural heritages* introduced Fall 1995.
Access to American Cultural Heritages: A Study Guide to Issues of Representation,
Delivery and Control, with J. Woo. Instructional Development Grant. 1997-98.

Doctoral dissertations directed

Snunith Shoham. *Organizational adaptation to the environment: The case of the public
library*. 1982.
Susan K. Martin. *Governance issues for automated library networks: The impact of and
implications for large research libraries*. 1983.
Joanne R Euster. *The activities and effectiveness of the academic library director in the*

environmental context. 1986.

Stanton F. Biddle. *The planning function in the management of university libraries: Survey, analysis, conclusions and recommendations*. 1988.

Doris Florian. *Information retrieval systems: Eine systematische Analyse der Probleme und Prioritäten für zukunftsweisende Lösungskonzepte: Von Expertise bis Artificial Intelligence*. [Information retrieval systems: A systematic analysis of problems and priorities for future-oriented concepts for improvement: From expertise to Artificial Intelligence.] Graz University of Technology, Austria, 1990.

John N. Gathegi. *Policy in the creation of scientific and technological information in developing countries: The case of agricultural information in Kenya*. 1990.

Samia Benidir. *Information seeking behavior during the decision making process: A case study*. 1991.

Annette Melville. *Managing with less: Resource strategies of university libraries*. 1994.

Michael G. Berger. *Information-seeking in the online bibliographic system: An exploratory study*. 1994. (UMI AAI9504745).

Ann M. Hotta. *Children, books, and children's Bunko: A study of an art world in the Japanese context*. 1995.

Richard Lemberg. *A life-cycle cost analysis of the creation, storage and dissemination of a digitized document collection*. 1995.

Ziming Liu. *Translation as a special channel of transborder information flow: An integrative study*. 1996.

Current and recent research

Searching, Selection, and Digital Libraries

Search support for unfamiliar metadata vocabularies. (DARPA \$954,184, 7/97-6/00.)

Seamless Searching of Numeric and Textual Resources. (IMLS Oct 1999 - Sept 2002).

Translingual Information Management. (DARPA).

Online Access in Multiple Database Environments. (US Department of Education, HEA IIA. \$48,000, 1994-96).

Ricoh University Grant. (Ricoh Silicon Valley, Inc., unrestricted research support, \$10,000, 1997; \$10,000, 1998).

History of Information Management

Emanuel Goldberg (Moscow 1881-Tel Aviv 1970), pioneer of electronic retrieval.

Biography. (Beta Phi Mu Harold Lancour Scholarship for Foreign Study, 1995, \$1,000).

Robert Gitler and the Japan Library School. Oral history, assisted autobiography, article. (Council on Library Resources, \$2,500).

Support for a Conference on the History and Heritage of Science Information Systems. (National Science Foundation, \$48,815, to the Chemical Heritage Foundation, Philadelphia, Arnold Thrackray, P.I., M. Buckland Co-P.I., 1998).

Information, Society, and Cultural Heritage

Cultural Property and Heritage Interpretation in the European Community. (University of California. Center for German and European Studies. Universitywide competition, \$3,000, 1996).

American Cultures Summer Faculty Fellowship, 1994.

Selected Service and Professional Activities

Advisor on educational programs in information and/or library studies:

University of Klagenfurt (Austria), 1972

Louisiana State University, 1982

University of Michigan, 1983

University of Minnesota, 1983

Rutgers University, 1987

Monash University (Australia), 1988

University of Tennessee, Knoxville, 1993

Fachhochschule Anhalt, Koethen, Germany, 1998.

American Society for Information Science. President, 1998.

Chair, Technical Program Committee, ASIS Mid-Year Meeting, 1991.

Member, Technical Program Committee, ASIS Mid-Year Meeting, 1992.

Member, ASIS Conferences and Meetings Committee, 1991-94.

Chair, Special Interest Group on Foundations of Information Science, 1994. (Vice-Chair 1993).

Special Interest Groups Cabinet Steering Committee, 1994-96.

University of California, Berkeley.

Academic Senate Committee on Educational Policy, 1992-95.

Committee on Academic Planning & Resource Allocation, 1995-97.

Graduate Council Review Committee, School of Education, 1992-94.

Information Planning Group, 1993. (Report adopted as the mandate for the School of Information Management & Systems.)

Academic Planning Board Working Group on Student Services, 1994.

Series Advisor, Monographic series *New Directions in Information Management*, Greenwood Press, 1986-.

Editorial boards:

Academic Press Dictionary of Science and Technology, 1992-93.

Advances in Librarianship, 1991-98.

Dekker Encyclopedia of Library and Information Science, 1998-;

Documents Numériques (Paris), 1999-;

Encyclopedia of Physical Science and Technology, 1st-3rd eds, 1998-;

Information Processing and Management, 1974-

International Information and Library Review, 1990-

Library and Information Science Research, 1980-92.

Library Hi Tech, 1991-

Library Quarterly, 1990-

Solaris: Dossiers du Groupe Interuniversitaire de Recherche en Sciences de l'Information et de la Communication, (Rennes, France), 1994-

Advisor on university library services:

Simon Fraser University, Burnaby, B.C., Canada, 1991.

University of Auckland, Auckland, N.Z., 1993.

Indiana Cooperative Library Services Authority. Vice-President, President-Elect, 1974-75.

Principal Investigator, Project to develop bibliographical instruction for users of public libraries, 1979-80. (HEA funded, jointly with Oakland Public Library).

Co-Principal Investigator, BIJOU, Berkeley-IBM Joint Study adapting and implementing office automation in an academic environment, 1983-85.

Chair, Management Review of the Division of Library Automation, (Office of the President, University of California), 1983.

California Library Networking Taskforce (California State Library), 1986-88.

Reviewer for National Science Foundation, Social Sciences and Humanities Council of Canada, U.S. Department of Education, *Information Processing and Management*, *Journal of the American Society for Information Science*, *Library Quarterly*, National Geographic Society, etc.

California Library Association.

Leadership Development Program Task Force, 1993-94.

Continuing Education Committee, 1996-97.

California Academic & Research Libraries. Planning Committee, 3rd Annual Conference, 1995.

Information Science Pioneers Project Advisory Board, University of South Carolina, 1995-

Co-editor, Special issue on the History of Information Science, *Journal of the American Society for Information Science*, April and September issues, 1997.

Planning committee, Conference on the History and Heritage of Science Information Systems, Oct. 23-25, 1998, Pittsburgh, PA.

University of Paris I - Panthéon-Sorbonne, Habilitation jury, 1998. (Formal examination of candidate for appointment to a professorial chair).

Honors, etc.

Who's Who in America.

University of Lancaster Library Research Unit awarded the Robinson Medal by the British Library Association for its computer-based library management games, 1972.

Visiting Scholar, Western Michigan University, 1979.

Guest, Visitors of Distinction Program, Universities Council of British Columbia, 1982.

Lazerow Lecturer, Florida State University, 1987; Indiana University, 1990.

Speaker, Distinguished Seminar Series, OCLC, 1987.

Visiting Fellow, Chisholm Institute of Technology, Melbourne, Australia, 1988.

Fulbright Research Scholar, Graz, Austria, 1989.

American Society for Information Science. SIG Member of the Year, 1994.

Lecturer, National Diet Library, Tokyo, 1996.

4th most-cited author among 411 faculty of schools of librarianship (*Journal of*

Education for Librarianship 23 (1983):161).
8th most-cited author in four leading Library and Information Science journals 1971-90.
(*Journal of Documentation*, 48 (1992):119).
33rd of 120 most-cited authors in Information Science, 1972-1995. (*JASIS* 49 (1998):
327-355).

Books

Robert Gitler and the Japan Library School. Lanham, MD: Scarecrow Press, 1999.
Assisted autobiography.

Redesigning Library Services. Chicago: American Library Association, 1992. Japanese
ed., 1994; Hungarian ed., 1998; Korean ed., 1998.

Information and Information Systems. New York: Greenwood Press, 1991. Paperback,
Praeger, 1991. Chinese ed., 1993.

Library Services in Theory and Context. New York: Pergamon, 1983; 2nd ed., 1988;
Japanese ed., 1990; Chinese ed., 1994; Croatian ed. in preparation.

Historical Studies in Information Science, edited by Trudi Bellardo Hahn and Michael
Buckland. Medford, NJ: Information Today, 1998.

Reader in Operations Research for Libraries. Englewood, CO: Information Handling
Services, 1976. Co-editor and contributor.

Book Availability and the Library User. New York: Pergamon, 1975.

Emanuel Goldberg and his Knowledge Machine. In draft.

2. Robert A. McCaughey

Anne Whitney Olin Professor of History, Barnard College, and Graduate Faculties,
Columbia University

Director of Barnard Electronic Archive and Teaching Lab (BEATL)

Recent Administrative Positions

Chairman of History Department, Barnard College, 1995-1998

Director of the Andrew W. Mellon Teaching Technologies Grant and Barnard Electronic
Archive and Teaching Laboratory [BEATL], 1997-2000

Vice President for Academic Affairs & Dean of the Faculty, Barnard College, 1987-1994

Director of the Mellon Curricular Consolidation Grant, 1993-1997

Founding Director of the Barnard First-Year Seminar Program, 1983-1987

Chairman of History Department, 1983-1987

Academic Degrees

Harvard University, Ph.D., History, 1970

University of North Carolina, M.A., History and American Studies, 1965

University of Rochester, A.B., History, 1961

Professional Recognitions

Gilder Lehrman Institute Distinguished Scholar Fellow, New York Historical Society,
Fall 1998

NEH Summer Fellow, American Maritime History Program, Mystic Seaport, 1996

Chair of Bancroft History Prize Committee, Columbia University, 1986, 1994

Emily Gregory Teaching Excellence Award, Barnard College, 1987

Elected Member of Society of American Historians, 1986

John Simon Guggenheim Fellow, 1975-76

American Council of Learned Societies Fellow, 1975-76

Charles A. Warren Fellow, Harvard University, 1972-73

Current Teaching Interests

The history of American colleges and universities

American maritime history and early maritime culture

The social history of American intellectual life

The uses of electronic networking in undergraduate teaching

Current Research and Writing

Writing a one-volume History of Columbia University, 1754-2000, to be published in 2003 by Columbia University Press as part of the University's 250th Anniversary in 2004.

Early research on a maritime history of New York in the Age of Sail

Principal Publications

Books and Monographs

Scholars & Teachers: Faculties of Select Liberal Arts Colleges and Their Place in American Higher Learning. Barnard College and the Mellon Foundation, 1995.

The American Nation: A History of the United States [with John A. Garraty]. 6th edition, Harper & Row, 1986.

International Studies and Academic Enterprise: A Chapter in the Enclosure of American Learning. Columbia University Press, 1984.

"The Transformation of American Academic Life; Harvard University, 1821-1892," *Perspectives in American History*, VIII (1974), pp. 239-332.

Josiah Quincy: The Last Federalist, 1772-1864. Harvard University Press, 1974.

Journal Articles

"But Can They Teach? In Praise of College Professors Who Publish," *Teachers College Record*, 95 (Winter 1993), 242-257.

"International Studies and General Education: The Alliance Yet to Be," *Liberal Education*, 70 (1985), pp. 343-374.

"'In the Land of the Blind': Non-Academic International Studies in the 1930s," *The Annals of the Academy of Political and Social Science*, 449 (1980), pp. 381-399.

"The Current State of International Studies in American Universities: Special Consideration Reconsidered," *Journal of Higher Education*, 51 (1980), pp. 381-390.

"Four Academic Ambassadors: International Studies and the American University Before the Second World War," *Perspectives in American History*, XII (1979), pp. 563-607.

"American University Teachers and Opposition to the Vietnam War: A Reconsideration," *Minerva*, XIV (1976), pp. 307-329.

"From Town to City: Boston in the 1820s," *Political Science Quarterly*, 88 (1973), 191-213.

"The Usable Past: The Harvard Rebellion of 1834," *William & Mary Law Review*, II (1970), pp. 587-610.

Book Reviews, Short Contributions, Journalism

In Academe, *American Historical Review*, *Change*, *Chronicle of Higher Education*, *Dictionary of American Biography*, *Journal of American History*, *Journal of*

Interdisciplinary History, [London] *Times Literary Supplement*, *Minerva*, *The New York Times*, *Political Science Quarterly*, *Teachers College Record*

Other Recent Writing

Review of Amherst College Faculty, *Teaching What We Do*, in *Teachers College Record* (Spring, 1994), pp. 421-423.

Featured review of Jaroslav Pelikan, *The Idea of the University: A Reexamination*, *The American Historical Review*, 98 (February 1993), pp.118-119.

Review of Joseph Brent, *Charles Sanders Peirce: A Life*, *The New York Times Book Review* (February 7, 1993), p. 11.

"Why Research and Teaching Can Coexist," *The Chronicle of Higher Education* (August 5, 1992), backpage.

Biographical entries for "Frederick A. P. Barnard" and "Josiah Quincy," in *American National Biography* (Oxford University Press, 1999).

Recent Public Addresses

"Orders of Learning," 1997 Barnard Honors Assembly, May 1, 1997.

"The Middle Period of Columbia's History: Barchester Towers in Gotham," to the Columbia University Seminar on Higher Education, February 24, 1996.

"Why Bother? The Institutional Case for College Professors Who Publish," to Corporate and Foundation Officers Association of the Northeast, New York City, January 4, 1994.

"The Henry Street Settlement, 100 Years On," for the centennial Observance of the Henry Street Settlement and the Visiting Nurses Association, New York City, April 7, 1993.

"Scribblers and Abstainers: Varieties of Professorial Life," AAHE Conference on Faculty Priorities, San Antonio, Texas, January 30, 1993.

Consulting Activities

On-Site Review of the Center for Educational Technology, Middlebury College, August 1998, for the Andrew W. Mellon Foundation.

On-Site Review of the sea and shore components the the "Semester at Sea" Program of the Sea Education Association, Wood's Hole, Massachusetts, March - July, 1998.

Advisory Committee, The G. W. Blunt White Library of American Maritime History, Mystic Seaport, May 1998.

Chair of Review of Middle School Social Studies Curriculum, The Dalton School, 1995-96.

Consultant for the International Division of the Ford Foundation, 1974-1978.

3. John M. Unsworth

Dean and Professor, Graduate School of Library and Information Science (and Professor, Department of English), University of Illinois at Urbana-Champaign, beginning August 15, 2003.

Education

- University of Virginia: Ph.D. in English, 1988
- Boston University: M.A. in English, 1982
- Amherst College: B.A. *Magna Cum Laude* in English, 1981

Recent and Forthcoming Publications

- *A Companion to Digital Humanities*, co-edited with Susan Schreibman and Ray Siemens. Under contract with Blackwell's. Due to be published in 2003.
- *Electronic Textual Editing*, co-edited with Lou Burnard and Katherine O'Brien O'Keefe. Co-sponsored by the Modern Language Association's Committee on Scholarly Editions and the Text Encoding Initiative Consortium, funded by a grant from the Andrew W. Mellon Foundation. Due to be published in 2004.
- "What is Humanities Computing, and What is Not?" in *Jahrbuch für Computerphilologie 4*, Georg Braungart, Karl Eibl & Fotis Jannidis, eds. Paderborn: mentis 2002.
- "Launching a scholarly electronic imprint," *Logos* 13.1 (2002): 43-48.
- "The Importance of Failure," in *The Journal of Electronic Publishing*, 3.2 (December, 1997).
- "Networked Scholarship: The Effects of Advanced Technology on Research in the Humanities," in *Gateways to Knowledge*, ed. Larry Dowler. MIT Press, 1997.
- "Electronic Scholarship" in *The Literary Text in the Digital Age*, ed. Richard Finneran. University of Michigan Press, 1996.
- "Living Inside the (Operating) System," in *Computer Networking and Scholarship in the 21st-Century University*, ed. Teresa Harrison and Timothy D. Stephen. SUNY Press, 1996.

Recent Teaching

- "Is Humanities Computing an Academic Discipline?" A seminar funded by the College of Arts and Sciences. This seminar led to a proposal for an MA in digital humanities, and to an NEH-funded seminar on the digital humanities curriculum. The MA has been approved by the State Council on Higher Education in Virginia, but is currently on hold because of the budget crisis.
- ENTC 312: 20th-Century American Literature (Bestsellers), Fall 2002

Editing and Curating

- Co-Curator (with Lynda Clendenning), “Rave Reviews: Bestselling Fiction in America,” an exhibition in Special Collections at the University of Virginia, February 22nd to June 10, 2002.
- Member, Editorial Board, *Arts and Humanities in Higher Education: an international journal of theory, research and practice*, 2001-
- Commissioning Editor, *Computers and the Humanities*, 1997-present
- Member, Editorial Board, *Journal of Electronic Publishing*, University of Michigan Press.
- Co-editor, *Research Reports of the Institute for Advanced Technology in the Humanities*
 - First Series (1993)
 - Second Series (1994)
 - Third Series (1995)
 - Fourth Series (1996)
 - Fifth Series (1997)
 - Sixth Series (1998)
 - Seventh Series (1999)
- Co-founder and Editor Emeritus, Member, Editorial Board, *Postmodern Culture: an electronic journal of interdisciplinary criticism* (published by Johns Hopkins University Press): issue editor for issues 1.1, 1.3, 2.1, 2.3, 3.3, 4.2, 5.1, 5.3.
- Member, Blake Archive Advisory Board, 1998-present
- Member, Dickinson Editorial Collective Advisory Board, 1998-present
- Member, Electronic Melville Committee, Melville Society, 1998-present
- Member, Multimedia Dante Project Advisory Board, Princeton University, 1998-present
- Member, Romantic Circles Advisory Board, 1997-present

Recent Grants

- Planning grant for Virtual Collections in American Studies, funded by the Andrew W. Mellon Foundation. \$46,829 (2002). Approved.
- Program Officer’s grant in support of *Electronic Textual Editing* (a volume co-sponsored by the MLA’s Committee on Scholarly Editions and the Text Encoding Initiative Consortium), submitted through the Modern Language Association and funded by the Andrew W. Mellon Foundation. \$40,250 (2002). Approved.
- “An Electronic Imprint at the University Press of Virginia,” a proposal to the Andrew W. Mellon Foundation in support of publishing originally digital scholarship, in partnership with the University of Virginia Press. \$640,000 (2001-2003). Approved.
- “The Walt Whitman Hypertext Archive,” an April 2000 Collaborative Research Award from the National Endowment for the Humanities (Project Director: Ed Folsom). \$150,000 (2000-2001). Approved.

- “The William Blake Archive,” an April 2000 Preservation and Access Award from the National Endowment for the Humanities (Project Directors: Morris Eaves, Robert Essick, Joe Viscomi). \$233,824 (2000-2001). Approved.
- Supporting Digital Scholarship, a proposal to the Andrew W. Mellon Foundation in support of IATH research and digital library integration, in partnership with the University of Virginia Library. \$1,000,000 (2000-2002). Approved.