

**INFORMATION ABOUT PRINCIPAL INVESTIGATORS/PROJECT DIRECTORS(PI/PD) and
co-PRINCIPAL INVESTIGATORS/co-PROJECT DIRECTORS**

Submit only ONE copy of this form for each PI/PD and co-PI/PD identified on the proposal. The form(s) should be attached to the original proposal as specified in GPG Section II.B. **DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPRISE THE CONFIDENTIALITY OF THE INFORMATION.**

PI/PD Name: Gregory B Newby

Gender: Male Female
Ethnicity: (Choose one response) Hispanic or Latino Not Hispanic or Latino

Race:
(Select one or more)
 American Indian or Alaska Native
 Asian
 Black or African American
 Native Hawaiian or Other Pacific Islander
 White

Disability Status:
(Select one or more)
 Hearing Impairment
 Visual Impairment
 Mobility/Orthopedic Impairment
 Other _____
 None

Citizenship: (Choose one) U.S. Citizen Permanent Resident Other non-U.S. Citizen

Check here if you do not wish to provide any or all of the above information (excluding PI/PD name):

REQUIRED: Check here if you are currently serving (or have previously served) as a PI, co-PI or PD on any federally funded project

Ethnicity Definition:

Hispanic or Latino. A person of Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race.

Race Definitions:

American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.

Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.

Black or African American. A person having origins in any of the black racial groups of Africa.

Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

WHY THIS INFORMATION IS BEING REQUESTED:

The Federal Government has a continuing commitment to monitor the operation of its review and award processes to identify and address any inequities based on gender, race, ethnicity, or disability of its proposed PIs/PDs. To gather information needed for this important tasks, the proposer should submit a single copy of this form for each identified PI/PD with each proposal. Submission of the requested information is voluntary and is not a precondition of award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information received from others. Any individual not wishing to submit some or all the information should check the box provided for this purpose. (The exceptions are the PI/PD name and the information about prior Federal support, the last question above.)

Collection of this information is authorized by the NSF Act of 1950, as amended, 42 U.S.C. 1861, et seq. Demographic data allows NSF to gauge whether our programs and other opportunities in science and technology are fairly reaching and benefiting everyone regardless of demographic category; to ensure that those in under-represented groups have the same knowledge of and access to programs and other research and educational opportunities; and to assess involvement of international investigators in work supported by NSF. The information may be disclosed to government contractors, experts, volunteers and researchers to complete assigned work; and to other government agencies in order to coordinate and assess programs. The information may be added to the Reviewer file and used to select potential candidates to serve as peer reviewers or advisory committee members. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records", 63 Federal Register 267 (January 5, 1998), and NSF-51, "Reviewer/Proposal File and Associated Records", 63 Federal Register 268 (January 5, 1998).

List of Suggested Reviewers or Reviewers Not To Include (optional)

SUGGESTED REVIEWERS:

Kichoon Yang, U. Texas Pan-American. Dept. of Mathematics.

Ray Larson, U. California @ Berkeley. School of Information Management and Systems.

David Dubin, U. Illinois at Urbana-Champaign. Graduate School of Information and Library Science.

Mark Rorvig. U. North Texas. School of Library and Information Science.

REVIEWERS NOT TO INCLUDE:

Bruce Schatz, U. Illinois. He has a reputation of not being a fair reviewer of other people's work.

CERTIFICATION PAGE

Certification for Principal Investigators and Co-Principal Investigators:

I certify to the best of my knowledge that:

- (1) the statements herein (excluding scientific hypotheses and scientific opinions) are true and complete, and
- (2) the text and graphics herein as well as any accompanying publications or other documents, unless otherwise indicated, are the original work of the signatories or individuals working under their supervision. I agree to accept responsibility for the scientific conduct of the project and to provide the required progress reports if an award is made as a result of this application.

I understand that the willful provision of false information or concealing a material fact in this proposal or any other communication submitted to NSF is a criminal offense (U.S.Code, Title 18, Section 1001).

Name (Typed)	Signature	Social Security No.*	Date
PI/PD Gregory B Newby		*ON FASTLANE SUBMISSIONS* SSNs are confidential and are not displayed	
Co-PI/PD			
Co-PI/PD			
Co-PI/PD			
Co-PI/PD			
Co-PI/PD			

Certification for Authorized Organizational Representative or Individual Applicant:

By signing and submitting this proposal, the individual applicant or the authorized official of the applicant institution is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding Federal debt status, debarment and suspension, drug-free workplace, and lobbying activities (see below), as set forth in Grant Proposal Guide (GPG), NSF 00-2. Willful provision of false information in this application and its supporting documents or in reports required under an ensuring award is a criminal offense (U. S. Code, Title 18, Section 1001).

In addition, if the applicant institution employs more than fifty persons, the authorized official of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of Grant Policy Manual Section 510; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution's expenditure of any funds under the award, in accordance with the institution's conflict of interest policy. Conflict which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

Debt and Debarment Certifications

(If answer "yes" to either, please provide explanation.)

Is the organization delinquent on any Federal debt?

Yes

No

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency?

Yes

No

Certification Regarding Lobbying

This certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding \$100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding \$150,000.

Certification for Contracts, Grants, Loans and Cooperative Agreements

The undersigned certifies, to the best of his or her knowledge and belief, that:

(1) No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.

(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, "Disclosure Form to Report Lobbying," in accordance with its instructions.

(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than \$10,000 and not more than \$100,000 for each such failure.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE	SIGNATURE	DATE
NAME/TITLE (TYPED)		02/12/00
TELEPHONE NUMBER	ELECTRONIC MAIL ADDRESS	FAX NUMBER

*SUBMISSION OF SOCIAL SECURITY NUMBERS IS VOLUNTARY AND WILL NOT AFFECT THE ORGANIZATION'S ELIGIBILITY FOR AN AWARD. HOWEVER, THEY ARE AN INTEGRAL PART OF THE INFORMATION SYSTEM AND ASSIST IN PROCESSING THE PROPOSAL. SSN SOLICITED UNDER NSF ACT OF 1950, AS AMENDED.

ITR/SW: TeraScale Retrieval Project Summary

Information retrieval is not a solved problem. In spite of the success of Web search engines at bringing fast and reasonably useful results to end-users, there is considerable work to be done to make Web searches more effective. Similarly, our ability to find information we need from full text data collections ranging from digital libraries to commercial database services such as Lexis/Nexis is limited.

The process of searching a collection of texts for items that match an information need is called information retrieval (IR). TeraScale Retrieval is intended to improve information retrieval, especially for relatively unstructured natural language text such as Web content.

One of the biggest factors missing from modern large-scale information retrieval systems is an ability to contribute to scientific knowledge about information retrieval. Large Web search services (AltaVista, Lycos, etc.) have not generally made their methods and processes known, and do not give away their software source code. In addition, the economic imperative for the companies that run these search services do not allow for significant experimentation in alternate methods, in order to advance knowledge about factors impacting retrieval performance.

The goals for this project are:

1. To enable scientific experimentation on IR systems. Scientific experimentation involves setting up variables and conditions. In order to carry out the experiments necessary to answer fundamental questions about IR, new software development is needed. A retrieval software toolkit will be authored and distributed. Some retrieval software is available in source form (such as ht://dig, SMART, WAIS, and others), but these allow relatively small numbers of alternate retrieval methods and are not geared toward software reuse and experimentation.
2. To focus on large-scale performance. Information scientists' contributions to modern information problems are limited by their experimental systems. To address real-world problems, experimental systems must be capable of high-speed performance with terascale datasets (hundreds of millions of documents with terabytes of raw data, millions of unique terms, multiple languages, and potential for quadrillions [petascale] of sub-documents or document fragments).
3. To share software and results. The TeraScale Retrieval project will produce a retrieval software toolkit emphasizing modular design, shared high-performance data structures and algorithms, and high-quality documentation. This toolkit will facilitate rapid development and deployment of the infrastructure needed for IR experimentation or production systems.

These goals are intended to allow the PI and his colleagues to address fundamental problems in IR by performing experiments, user-based evaluation and real-world deployment studies. The project will include development of large-scale retrieval systems with an emphasis on scientific discovery. Rather than moving rapidly from academic use to commercialization (such as Yahoo!, Google and Lycos), the TeraScale Retrieval project will retain a focus on experimentation and evaluation to contribute to scientific knowledge about information seeking and use.

TABLE OF CONTENTS

For font size and page formatting specifications, see GPG section II.C.

Section	Total No. of Pages in Section	Page No.* (Optional)*
Cover Sheet (NSF Form 1207 - Submit Page 2 with original proposal only)		
A Project Summary (not to exceed 1 page)	1	_____
B Table of Contents (NSF Form 1359)	1	_____
C Project Description (including Results from Prior NSF Support) (not to exceed 15 pages) (Exceed only if allowed by a specific program announcement/solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)	15	_____
D References Cited	2	_____
E Biographical Sketches (Not to exceed 2 pages each)	2	_____
F Budget (NSF Form 1030, including up to 3 pages of budget justification)	6	_____
G Current and Pending Support (NSF Form 1239)	1	_____
H Facilities, Equipment and Other Resources (NSF Form 1363)	1	_____
I Special Information/Supplementary Documentation	2	_____
J Appendix (List below.) (Include only if allowed by a specific program announcement/ solicitation or if approved in advance by the appropriate NSF Assistant Director or designee)	_____	_____
Appendix Items:		

*Proposers may select any numbering mechanism for the proposal, however, the entire proposal must be paginated. Complete both columns only if the proposal is numbered consecutively.

Project Description for NSF 99-167

ITR/SW: TeraScale Retrieval

1. Introduction

Information retrieval is not a solved problem. In spite of the success of Web search engines at bringing fast and reasonably useful results to end-users, there is considerable work to be done to make Web searches more effective. Similarly, our ability to find information we need from full text data collections ranging from digital libraries to commercial database services such as Lexis/Nexis is limited.

The process of searching a collection of texts for items that match an information need is called information retrieval (IR). The TeraScale Retrieval project is intended to improve information retrieval, especially for relatively unstructured natural language texts such as Web content.

One of the biggest factors missing from modern large-scale information retrieval systems is an ability to contribute to scientific knowledge about information retrieval. Large Web search services (AltaVista, Lycos, etc.) have not generally made their methods and processes known, have not given away their software source code, and have not published retrieval performance studies. In addition, the economic imperative for the companies that run these search services do not allow for significant experimentation in alternate methods, in order to advance knowledge about factors impacting retrieval performance.

There is a substantial amount of work on furthering scientific knowledge about IR that happens mostly outside of the commercial domain. Information scientists are scholars who seek to improve the performance of IR systems through development and testing of systems, modeling, or user-based evaluation. One of the most important forums for evaluating new ideas for IR systems is the annual Text REtrieval Conference (TREC, see Voorhees and Harman, 2000). TREC brings together scholars and industry representatives to work on IR challenges.

Industry participants at TREC typically do not share their particular methods or software (recent industry participants include Apple, IBM, Microsoft, Claritech and Lexis/Nexis), but they may share algorithms or equations used for some particular functions in their software. By contrast, the academic participants in TREC tend to offer relatively full disclosure on how their systems work, although relatively few actually make their source code available to others.

There are several areas that are not addressed by TREC, and these are the main themes to be addressed by the TeraScale Retrieval project.

1. **Scaling.** Most academic systems have difficulty with real-world datasets (e.g., tens of millions of unique Web documents). TREC has mostly used a 2GB dataset of news articles, which is only a fraction of the size indexed by a typical Web search engine. In 1998 and 1999, a small number of TREC participants started work with a larger 100GB dataset of Web documents, which is a step in the right direction.
2. **Speed.** While Web search engines are able to respond to many requests per second, academic systems are usually not as fast. For the large Web corpus used in the 1999

TREC, only one system (the PI's) performed searches in substantially less than one second each. Most of the other systems' search times increased approximately linearly with the dataset size, so that a .1 second search with a 2GB dataset would take 5 seconds with a 100GB dataset.

3. Accessible software. Although some groups at TREC make their IR system software available, most do not. There are relatively few modern IR systems with fully-accessible software (ht://dig is one of the best examples, but it is intended for production use rather than evaluation experiments).
4. Emphasis on performance. For TREC participants, performance is focused on recall and precision (as discussed below). Inefficient data structures, disproportionately large index files, excessive memory usage, etc. are not a factor, often resulting in systems for TREC that are limited in their practical usefulness. By contrast, Web search engines tend to emphasize speed over other factors, so that IR techniques proven effective in TREC and elsewhere are avoided in favor of techniques that can generate results very quickly.
5. Modern and modular software design. Historically, experimental IR was carried out by building "toy" systems suitable for running a particular study. TREC has, to some extent, continued that pattern, with a proliferation of systems that are tuned for one particular type of retrieval model. Software re-use, like software accessibility, is typically not emphasized. Systems tend to be fairly fragile and non-modular. Modern software design approaches (notably object oriented design and implementation) are less common than old-style monolithic programming. Modern toolkits and packages, such as the Standard Template Library (for C++ programming) or Purify (a product for seeking memory leaks in programs) have not been widely adopted by information scientists.

Many valid purposes are served by TREC, by Web search engines, and by other experimental work in IR. The purpose of this proposal is to address specific areas that have been identified as key for advancing knowledge about information searching and retrieval, but are unlikely to be addressed by most current efforts.

The goals and research questions for this project are:

1. To enable scientific experimentation on IR systems. Scientific experimentation involves setting up variables and conditions. In order to carry out the experiments necessary to answer fundamental questions about IR (as discussed below), new software development is needed to manipulate variables and set conditions for controlled experiments.
2. To focus on large-scale performance. Information scientists' contributions to modern information problems are limited by their experimental systems. In order to address real-world problems, experimental systems must be capable of high-speed performance with terascale datasets (tens or hundreds of millions of documents, many millions of unique terms, multiple languages, and potential for quadrillions [petascale] of sub-documents or document fragments).
3. To share software and results. The TeraScale Retrieval project will produce a retrieval software toolkit emphasizing modular design, shared high-performance data structures

and algorithms, and high-quality documentation. This toolkit will facilitate rapid development and deployment of the infrastructure needed for IR experimentation or production systems.

These goals are intended to allow the PI and his colleagues to address fundamental problems in IR by performing experiments, user-based evaluation and real-world deployment studies. The project will include development of a full-scale retrieval system, but with an emphasis on scientific discovery rather than production use. For example, instead of moving rapidly from academic use towards commercialization (such as occurred with Yahoo!, Google and Lycos), the TeraScale Retrieval project will retain a focus on experimentation and evaluation.

2. The Need for TeraScale Retrieval

This section addresses areas that support the need for the TeraScale Retrieval project. There are many contemporary efforts to address challenges of the Information Age, but few, if any, that are prepared to deliver the emphasis of TeraScale Retrieval.

2.1. The Need: Accessibility to information has outpaced our ability to effectively manage information

In 1999, the U.S. Presidential Information Technology Advisory Committee (PITAC, 1999) released a report intended to chart the near-term future of high-technology research and research funding. This report partially shaped the NSF's ITR program. PITAC identified several themes or areas that TeraScale Retrieval will address:

- "Transforming the way we deal with information ... requires significant improvements in data access methods, including high performance information systems and tools to help individuals locate information and present, integrate, and transform the information in meaningful ways (p. 13)."
- The PITAC report urged the development of "software for managing large amounts of information, and fundamental research in capturing, managing, analyzing and explaining information and making it available for its myriad of uses (p. 4)."
- The report identified the role of information retrieval for research and development in organizations (p. 16), for supporting electronic commerce (p. 13), and for managing information (p. 40).

These points from the PITAC report are not being addressed by TREC or by for-profit Web search engines or database vendors. To develop systems to effectively search and utilize large amounts of information, a rigorous scientific approach to exploring alternate IR methods is called for. As described above, the supporting software and infrastructure to support this type of scientific exploration does not yet exist.

One of the most telling sections in the PITAC report is found in a sidebar on page 6:

The PITAC members from industry were unanimous in their opinion that it is not feasible for the private sector to assume responsibility for long-term, high-risk research, in spite of the success of the information technology industry.

This statement is contrary to popular perceptions that current Web search engines and other technologies are (or will be) a panacea for information seeking. In fact, as PITAC recognized, search engine and database companies have little opportunity for experimentation, and seldom contribute meaningfully to furthering scientific knowledge about information seeking and use.

2.2. The Need: Growth of Available Information

In a well-known analysis, Wurman states, “a weekday edition of The New York Times contains more information than the average person was likely to read in a lifetime in seventeenth-century England (1990, p. 32).” Yet, the human capacity to deal with all this information is probably not much different than it was in the 1600s. Today, one of the biggest sources of growing information is the Internet. Although XML and other developments offer some promise for the Internet’s vast data store to become more searchable, few Internet users of today would argue that the Internet is anything other than a big mess.

The Web and its affiliated standards and protocols is in some ways the largest body of information in history. The Internet Archive estimates the growth of information accessible on the Internet by HTTP at 10% per month (<http://www.archive.org>), and have archived over 14TB of data.

Depending on how one chooses to measure size, the Internet may still be smaller than the US Library of Congress collection – but there is little doubt that the Internet is growing more quickly the number of books in the LoC. The Matrix Information and Directory Service, in cooperation with Network Wizards, has charted the growth of the Internet from 1969 to the present. Since the introduction of the Internet Protocol in 1989, the number of Internet-connected hosts in the DNS increased at a rate of 109% per year, but may have started to slow in late 1999 (Metcalf, 1999). In real numbers, Network Wizards estimates the number of Internet hosts in January 2000 at 72,398,092 (<http://www.isc.org/ds/WWW-200001/report.html>).

2.3. The Need: Information retrieval progress and practicalities

At about the same time the PITAC report was released, an NSF-sponsored invitational workshop on the future of information retrieval research was held (Korfhage et al., 1999). The workshop attempted to chart the future needs of IR in order to increase the contributions of information science in addressing the challenges of the Information Age. The report’s recommendations included the following:

- “To develop specifications for [and implement] a complete and modular set of IR tools to be made available to the IR community (p. 5).”
- To create an infrastructure for “sharing and distribution of IR tools (p. 5).”
- To overcome limitations of many retrieval research systems, viz., “they lack modularity and do not facilitate interactive and operational retrieval experimentation (p. 5).” This comment was directed particularly at the limitations of TREC.

Information scientists have developed many techniques that could be helpful to the challenges of information search and management that PITAC recognized, but relatively few are found outside

of IR laboratories or TREC. Examples of techniques that will be incorporated in the TeraScale Retrieval software toolkit include:

- Integration of user process models (Dervin and Nilan, 1986) or user histories (see the description of the TREC filtering track in Voorhees and Harman, 2000) to aid in identifying useful documents
- Enabling relevance feedback (Salton and McGill, 1983) by selecting a series of useful documents or document extracts (relevance feedback is partially implemented as “find more documents like this one” in some search engines).
- Incorporating process models. For example, some documents might be useful for people at the beginning of their information search, while other documents might be more useful for people who are at the end of their search (Kulthau, 1993).
- Addressing different types of information needs. Matching queries to documents has been the main activity of IR systems. Recently, search engines such as AskJeeves have been developed to address a different need: to get an answer to a question. Other types of information needs exist, and may be best met with different types of systems (Newby, 1998).
- Document characteristics such as the discourse structure (Liddy, 1997) and linguistic characteristics (Haas, 1996, 1997) may be used, along with semantic or syntactic analysis of the query (see the discussion of the question answering track in TREC by Voorhees and Harman, 2000), to improve the appropriateness of results.

The TeraScale Retrieval project will integrate existing methods from the IR literature so that effective science may be done to determine these and other methods’ usefulness. Basic experiments on, for example, the relationships between relevance feedback and a deep linguistic analysis of queries have simply not been done. This lack of basic research is, we believe, primarily a result of not having a high performance software toolkit that can be easily configured for retrieval experiments. The TeraScale Retrieval project will produce such a toolkit.

2.4. The Need: Performance scaling

As mentioned above, the growth of Internet hosts is estimated at over 100% per year. We may infer that the number of Internet users and quantity of available information is growing at about the same rate.

Web search engines have generally not been able to keep up with this growth. The largest, AltaVista, only indexes about 30% of the Internet’s content, and is only able to update its database a few times per year (Silverstein and Henzinger, 1999). Although the problem of indexing the whole Internet in a reasonable time is outside the scope of this proposal, this problem is central to a strong need addressed by TeraScale Retrieval: effective scaling.

A recent report advocated the possible wisdom of waiting to purchase computers for large-scale scientific research. This recommendation was based on Moore’s Law, which states that the number of transistors on state-of-the-art microprocessors doubles about every 18 months

(implying a proportional increase in computing capacity). Unless the growth of the Internet slows substantially, however, Moore's Law does not help with the challenges of large-scale IR: the Internet is growing faster than microprocessors (and other computer components) can catch up.

Furthermore, microprocessor computing capacity is only one factor in computer performance. For example, while microprocessor speed from 1991 – 2000 increased from about 20Mhz to 800Mhz (400-fold), the typical amount of memory in a mainstream computer increased from 4MB to 64MB (16-fold), and the typical hard disk increased from 40MB to 4GB (100-fold). During the same period, the Internet grew from fewer than ½ million hosts to over 70 million (perhaps 140-fold). Thus, while Moore's Law gives us hope that our computational capabilities will continue to increase, it does not offer any hope that we will be able to better handle the Internet's information without improving IR methods.

There are several techniques for addressing scaling issues. TeraScale Retrieval will investigate the following:

- What indexing methods will help perform searches in better-than-linear time as the size of the database grows? As the dataset grows from 2GB (TREC size) to 200GB (small search engine) to 2TB (larger search engine), it is important that query time does not increase proportionally.
- How can results from multiple smaller databases be merged? This challenge was addressed by WAIS and the Z39.50 protocol in the early 1990s, but today's systems favor a large monolithic collection. We believe that different collections may benefit from different retrieval approaches. More effective retrieval results may be obtained by searching a specific sub-collection (with appropriate IR methods for that collection), or by merging results from multiple collections. The Web meta-search engines have taken steps in this direction already.
- How can large datasets be segmented? A large corporate Web site might contain tens of thousands of Web pages – but perhaps this content could be automatically summarized into one meta-document. This approach may allow us to collapse millions of pages from tens of thousands of servers into a few tens of thousands of meta-documents, where each meta-document represents related content from a single source. This approach does not preclude searching a complete dataset of millions of pages, but may result in improved search performance while avoiding a more massive search.

3. The TeraScale Retrieval Project

This proposal is for a three-year project that will focus on software development and scientific comparison of IR methods and techniques. It will seek to make an IR software toolkit available to researchers to speed deployment of experimental systems and provide an accessible common basis for system comparison.

The software will be extensively tested in a large-scale retrieval environment. Terabyte-scale collections will be indexed and available for searching and retrieval. There will be no effort to

become the next “great Web search engine.” Instead, TeraScale Retrieval will focus on the science of real-world evaluation of search performance, retrieval effectiveness and scaling.

3.1. Timeline

Activities will be carried out in parallel. Existing systems, algorithms and data formats will be evaluated and incorporated or modified as needed. Throughout the project, source code will be made freely available.

Year 1 will emphasize the development of low-level modules and a common set of data formats for IR. Because most IR systems, at their core, require an inverted index and some sort of Boolean “OR” or “AND” process to determine relevant documents, it is reasonable to have a shared data format and set of core functions for different retrieval methods (as discussed below). For example, the PI has implemented a single data format specification that may be used for Boolean-style retrieval, vector space retrieval, and probabilistic retrieval.

Each module will be documented internally (source code comments) and externally (user guides). Performance evaluation will be completed for each module, in order that intelligent choices may be made about how to combine them. A choice of a standardized programming language (C++) and compiler (g++) will speed deployment to multiple platforms and computing environments (Linux, Solaris, AIX, IRIX, etc.).

By the end of year 1, a collection of high-performance modules for fundamental retrieval models will be available, enabling construction of a large variety of IR systems. Performance tradeoffs for different methods will be documented, and scaling issues from 2GB to 20GB to 200GB will be assessed. Experimentation on different performance characteristics for fundamental retrieval system models and techniques will be carried out, using TREC relevance judgments and other means as appropriate.

Year 2 will emphasize deployment and evaluation. Large-scale datasets will be built and made available for use. Performance evaluation will be carried out, and software development and deployment will continue. IR researchers will be invited to use the TeraScale Retrieval modules to integrate with their own systems, and donations of new or alternate modules will be sought.

Performance will be evaluated based on precision of results (discussed below), search speed and other system-oriented measures, and analysis of actual usage logs. Other researchers or projects may also engage in user-based evaluation, in which individuals are given particular search tasks and asked to evaluate system performance (funding for extensive user-based evaluation is not included as part of this proposal in order to emphasize development of software and infrastructure for experimentation, integrated with systems-based experimentation using test collections, as described further below).

By the end of year 2, dozens of different approaches to IR will be available through a unified software toolkit, enabling hundreds or thousands of possible system configurations. For example, a researcher might want to compare a system using the vector space model with a particular term stemming method and relevance feedback model with a system based on latent semantic indexing (Deerwester et al., 1990) using a different stemming method but the same

relevance feedback model. Scaling issues from 2GB to 2TB will be assessed, as will methods for collection segmentation and data fusion.

Year 3 will continue with software development and deployment, but will seek an expanded notion of the retrieval dataset. While years 1 and 2 will emphasize retrieval from Web documents, and will operate mostly with English, year 3 will seek further involvement with multi-lingual and multi-media documents. By year 3, developments in XML and other metadata standards are expected to provide a rich basis for developing new approaches to IR in the Internet environment.

Scaling and performance will continue to be of interest. Moore's Law indicates that multi-TB collections will be practical for the relatively small budget for this proposal. Research on collection segmentation and data fusion may lead away from monolithic databases towards federated collections.

3.2. Deliverables

TeraScale Retrieval will provide software and performance measures to the IR community, large-scale retrieval systems for public consumption (but without attempting to compete with existing search engines – instead focusing on real-world evaluation), and experimentation on fundamental issues in IR for multi-faceted systems.

Specific goals, based on a fall 2000 grant start date, include:

1. Fall 2000
 - Deploy 1TB file and Web server for retrieval dataset construction and IR evaluation.
 - Design, implement, document and release software and specifications for generic inverted file structures, term indexing modules, and other fundamental software.
 - Provide access methods (CVS, FTP, etc.) to the IR community to use and contribute to the software toolkit.
2. Spring 2001
 - Implement and evaluate the performance of common algorithms and data structures for their suitability in large-scale IR systems.
 - Provide a common experimental system suitable for Boolean, vector space, probabilistic, and latent semantic indexing styles of IR and variants (including different stemmers, pre- and post-processing of data, HTML parsers and spiders, etc.)
 - Design, implement, document and release software and specifications for useful algorithms and data structures. Provide for integration of these with "turnkey" IR systems built from the toolkit.
 - Perform experiments on different factors in the toolkit impacting retrieval performance, speed, database size, etc. and publish the results.

3. Fall 2001

- Integrate specific non-generic functions into the toolkit. Such functions are not representative of the major classes of IR systems (as described below), but are commonly used. Examples include variations on term weighting schemes, document parsing schemes, part of speech tagging, and semantic analysis of queries and documents.
- Integrate user models, relevance feedback, filtering, process models and other methods for customizing searchers based on characteristics of information need situations, user groups or individual users.
- Upgrade the file server to handle larger datasets and more traffic.
- Work on segmenting collections and techniques for special-purpose collections. Integrate these capabilities into the toolkit, and continue to experiment on their effectiveness.

4. Spring 2002

- Distribute ready-made retrieval system options in the toolkit. This will enable researchers to specify system characteristics to compile/build the program they choose, then expand/extend as desired. It will also facilitate the development of production systems based on the toolkit.
- Add toolkit components appropriate for the contemporary Web. Whereas support for various HTML tags and document structures common in the year 2000 will have been added previously, new anticipated developments for 2002 include an emphasis on XML and DTDs for specifying document structure and content.
- Integrate facilities for multiple views of data (e.g., Korfhage's [1997] VIBE; Kohonen et al.'s [1999] self-organizing maps, etc.).
- Continue experimentation and publication of results with an emphasis on contemporary data types.

5. Fall 2002

- Add toolkit components specific to data types not emphasized previously. These are expected to include:
 - a. Live retrieval: Identification of relevant "documents" from dynamic datasets such as television programming, commerce database inventories and electronic public debate.
 - b. Multimedia data: Techniques for retrieval of pictures, moving images, sounds, etc. are known but largely incompatible with techniques for retrieving

text. Integration of non-text methods to the toolkit will strengthen its utility, and enable further experimentation and comparison.

- c. Video browsing: Rapidly scanning through video data in a networked environment (e.g., Green et al., 2000).
 - d. Networked cooperative servers: We anticipate projects such as the Internet2 Distributed Storage Infrastructure (<http://dsi.internet2.edu/>), Web server developments, and metadata developments will result in an increased reliance on federated networked collections, rather than centralized monolithic collections. The TeraScale Retrieval project will be positioned to provide innovations for effective IR in these newly emerging environments.
- Deploy a new file and database server (estimated at 250-400% the capacity of the fall 2000 server) for expanded experimentation on scale and integration of multimedia data types.
 - Experiment on the integration of multimedia with text retrieval. What factors influence effectiveness? Is this a data fusion problem, or are there more fundamental differences?

6. Spring 2003:

- Insure the toolkit is exhaustively documented, consistently implemented and effectively packaged. Plan for continued development and deployment.
- Set an agenda for future IR development requirements and begin working on that agenda.
- Continue experimentation on new data types, larger-scale deployment, fusion of results, networked cooperation among servers, metadata, and other contemporary topics.

Reporting for this project, beyond that formally directed by the NSF (annual and final reports), will be primarily in the form of the deliverables listed above. The TeraScale Retrieval project will be highly public in nature, with ready access to “production” toolkit releases, as well as “development” versions. The success of Linux, Apache, GNU and other open source projects has indicated the likelihood of benefiting from sharing the software as widely as possible, and allowing interested parties to contribute their ideas and code to the development.

4. Retrieval Methods and Retrieval Evaluation

Modern computerized information retrieval has a 40-year history, starting with searching bibliographic records. The first IR systems used magnetic tape or punched card records of bibliographic data (author, title, and subject keywords for documents such as journal articles or books). Today, the emphasis has shifted to searching the full text of items such as Web pages, journal articles, encyclopedia entries, etc., but the basic approach has not changed very much.

The fundamental process in IR remains the matching of queries to documents (or to document surrogates such as bibliographic records).

A query is the manifestation of an information need. Although the information need may be viewed as a human cognitive phenomenon (see Ingwersen, 1992), IR systems do not have any sort of direct access to human cognitive structures or thoughts. Instead, the human information seeker must express her information need as a query so that it may be processed by the IR system. Historically, effective query generation required extensive knowledge of the IR system to be used, and thus was carried out by trained information intermediaries. Today, the information seekers themselves often attempt to formulate queries and submit them directly to the IR system (often with unintended or unsatisfactory results).

Although there are hundreds of variations on IR in commercial and experimental systems, there are a relatively small number of common types, with considerable overlap in their fundamental data structures and methods. These types are:

1. Boolean systems. Common examples are online library catalogs, where subject keywords are combined to retrieve matching book citations. AltaVista and other search engines offer Boolean capability, so that people can specify AND, OR and NOT combinations of terms.
2. Vector space systems (see Salton and McGill, 1983). Documents are represented as a vector in a multidimensional term space. Queries are also represented as vectors, so that similar (close) documents may be retrieved. Compared to Boolean systems, vector space enables more sophisticated term weights and can consider term importance within a particular document.
3. Probabilistic retrieval (for discussion see vanRijsbergen, 1979 or Frakes and Baeza-Yates, 1992). Determine the probability of whether a document is relevant to a query based on prior relevance judgments (which are sometimes simulated).
4. Latent semantic indexing (Deerwester et al., 1990). Analyze statistical relationships among terms and documents to generate a vector space where term vectors are not mutually unrelated (orthogonal), as with vector spaces.

For each of these main methods, and nearly all other IR systems, the fundamental process for retrieval consists of building an index that allows rapid lookup of documents containing query terms (or terms related to the query). Various schemes for assigning weights to query terms and terms within documents are applied, and mechanisms are employed to rank documents so documents that are more relevant are retrieved first. Although some IR systems use a traditional database system as the “back end” for storing term and document information, many use local files to decrease I/O latency.

IR effectiveness is most commonly measured using precision and recall. When an end user submits a query, she receives a list of documents that match that query. Precision is the proportion of these results that are relevant. Recall is the proportion of relevant documents from the entire dataset that were part of the list of documents.

High precision is often associated with low recall, and vice versa. Consider that a recipe for perfect recall (that is, all relevant documents in a dataset are part of a response to a query) is to retrieve all documents in the dataset. But a recipe for perfect precision might be to only retrieve a single document – anticipating that if this document is relevant, then 100% precision has been obtained. Information scientists attempt to balance recall versus precision, although contemporary search engines focus more on “early precision,” with the notion that most information seekers will not review more than one or two screens of results, regardless of how many relevant documents are in the dataset (Silverstein and Henzinger, 1999).

One of the biggest criticisms of the recall/precision approach to retrieval evaluation has to do with the problem of relevance. Instead of incorporating the context- and situation-dependent aspects of relevance with such factors as order effects, response set size and authoritativeness, recall and precision studies typically consider relevance as a “yes/no” decision that occurs without any particular user context (for a review of factors impacting relevance, see Schamber et al., 1991). The TeraScale Retrieval project supports the notion that different relevance criteria must be applied for different retrieval situations and users, and that different system configurations are needed to meet different criteria.

Another problem with assessing the recall and precision performance of an IR system is that relevance judgments are required. One of the primary stated goals of TREC is to create a collection of relevance judgments for a substantial number of queries (450 in 1999) and documents (about .5M). For much larger datasets, and greatly varied queries – as are found on the Web – such test collections are difficult or impossible to construct. The TeraScale Retrieval project will use the TREC test collection of relevance judgments when appropriate to evaluate retrieval performance.

Real-world evaluation of large-scale datasets typically emphasizes simple server log-based measures: in response to a query, how many documents are actually selected for evaluation (e.g., measured by clicking through a Web document hyperlink). Although this gives little data about the actual information need behind the query and whether the viewed document was really perceived as relevant, it does give a large quantity of potential evaluation data.

Information scientists often strive to evaluate retrieval system performance in terms of actual user needs and uses. Such studies might involve tracking longitudinal use of datasets by a population, focus groups, interviews, or extensive survey data – often coupled with analysis of system usage logs. Such studies, exemplified by Bishop et al. (2000), are often hampered by the lack of experimental IR software that could be tuned or modified on demand without a costly and time-consuming code re-write.

TeraScale Retrieval will provide facilities for all of these types of evaluation. Although the TREC test collection is smaller than large-scale Web collections envisioned, the software toolkit will enable rapid and complete experimentation on different retrieval methods on the TREC collection. This experimentation will provide greater control over variables than has been possible previously. This is because most systems at TREC can provide only a few different retrieval methods, and comparison across systems is problematic because of the great number of uncontrolled or unknown differences. Through this effort, a major recommendation of the Korfhage et al. (1999) report will be addressed.

Further evaluation enabled by the TeraScale Retrieval project will include click through studies on the use of search results, and comparison of different interface methods and techniques. Although human subjects research is not part of this proposal (other than analysis of anonymous log files or feedback forms), it is anticipated that the PI and his colleagues will use the software toolkit to solicit detailed evaluation from users with real information needs in related research.

5. Qualifications of the PI and Institution

The PI has been a participant in TREC since 1996 and an active developer of experimental IR systems and interfaces since his doctoral work in 1990. He was one of only seven participants in TREC's "very large corpus" (VLC) task in 1999, which consisted of running 10,000 queries against a 100GB collection of Web documents. By emphasizing system search speed and efficiency, and implementing a multi-way tree data structure, the PI's system performed the 10K queries in 52 seconds (this was 2 orders of magnitude faster than the next fastest system). Currently, the PI is integrating components to increase retrieval performance in the system, and further incorporating and testing generic IR modules.

The development of scalable IR systems for retrieval is the primary research area of the PI, but the TeraScale Retrieval project cannot happen without external support for graduate student programmers, researchers and documentation specialists, and equipment. Rather than limiting his IR development and evaluation to the relatively limited sphere of TREC and related activities, the PI seeks to contribute to real-world large-scale retrieval in ways that most information scientists and commercial systems have not.

6. Links to the Program Solicitation

The categories of NSF's Information Technology Research (ITR) are intentionally broad. This proposal identifies with the "software" category, but is intended to address other categories of research as well.

The program solicitation directs software development to be innovative, rather than extending existing toolkits. It is our contention that the lack of a toolkit for IR system experimentation has been a key detriment to information scientists' contributions to modern information problems. Although the proposed toolkit will integrate many well-known techniques (such as term weighting, vector space retrieval, etc.), it will do so in an innovative way by emphasizing software efficiency, reuse, and interoperability. Although such techniques are well known in the software industry, they have been notably lacking in experimental IR, as discussed above.

The issue of scalability in software implementation is an important concern for the TeraScale Retrieval project. In fact, scalability is considered the main barrier to deployment of potentially effective IR techniques in large-scale real-world systems. Due to the lack of contribution to fundamental research by commercial Web search engines (as observed in the PITAC report), there is virtually no evaluation literature on large-scale retrieval systems. (Note that large-scale database systems are specifically excluded here – in fact, large-scale databases are the topic of several academic conferences. However, IR systems tend to use different file structures, access methods and query processing methods than database systems, despite some overlap in purpose.)

The solicitation's emphasis on information management is well served by this proposal. It seeks, "fundamental research activities dealing either with online content (transforming the kind, quality or amount of material online) or with access (increasing the utility of online information via research on quality, economics, searching, or other related areas)." This project addresses fundamental research on IR, and will enable such research by the PI and others by implementing the necessary software infrastructure. Information management also includes multimedia, data fusion, multi-lingual systems, and other topics addressed in this proposal.

The advanced computational science and scalable information infrastructure themes are also relevant to this proposal. Although it is not the intent to develop radically new algorithms, it is expected that empirical large-scale performance evaluation of algorithms we implement will be of value to computational scientists. The IR evaluation software toolkit is proposed as a valuable part of the information infrastructure for the future. Contemporary advances in programming languages and techniques include object-oriented programming, generic programming, and toolkits such as the Standard Template Library. We envision a future where the information retrieval infrastructure we develop for this project will be as important as these advances for furthering IR system development and understanding.

Educational impact is an NSF criterion of strong interest to the PI and other academic faculty interested in IR. In addition to funding graduate assistantships for the duration of the project, TeraScale Retrieval is intended as a useful teaching tool. Today, teachers might use Cornell University's SMART software, [ht://dig](http://dig), or WAIS to allow students to perform small-scale retrieval experiments. The software toolkit this project will develop will enable far more sophisticated student experimentation, with less effort.

7. Conclusion

Science progresses by developing new ideas and testing them. Information scientists interested in information retrieval often seek scientific progress and knowledge by developing IR systems and testing them in various conditions. Today, progress in IR is hampered by several factors that the TeraScale Retrieval project seeks to overcome:

1. This project will widely disseminate high-quality software suitable for use in IR experimentation. Currently, very little IR software is freely available, and almost none is suitable for experimentation.
2. This project will carry out experimentation that transcends particular general retrieval approaches. Although forums such as TREC enable gross comparisons among different systems, there is not enough control of system differences to generate results with confidence.
3. This project will emphasize scalable solutions to retrieval problems. In spite of Moore's Law, the quantity of information has outpaced our ability to gather and utilize it. Furthermore, many experimental IR systems do not scale well beyond a few GB – while the search engines that do operate on a large scale have not generally made their software and methods available.

The PITAC report (PITAC, 1999) and the NSF-sponsored Invitational Workshop on Information Retrieval Tools (Korfhage et al., 1999) have identified an immediate need for software and evaluation studies that address the challenges of searching large quantities of electronic text. These challenges have not been met by TREC or by commercial database or search engine vendors. Relatively few information scientists have engaged in experimentation on large-scale retrieval (although some have, such as Witten et al., 1999). The result is that fundamental research on how different retrieval techniques might address modern information searching challenges is lacking.

The TeraScale Retrieval project will deliver high-quality, scalable and widely available software to the community of scholars and practitioners interested in information retrieval. It will further scientific knowledge of how different factors influence retrieval performance. This knowledge will assist in addressing the challenges of dealing with large quantities of relatively unstructured text in environments ranging from the Web, to digital libraries, to corporate intranets.

The PITAC report includes a main theme in NSF's ITR program solicitation:

“Transforming the way we deal with information ... requires significant improvements in data access methods, including high performance information systems and tools to help individuals locate information and present, integrate, and transform the information in meaningful ways (PITAC, 1999, p. 13).”

If funded, TeraScale Retrieval has the potential to greatly contribute to these important goals.

References Cited

- Bishop, Ann Peterson; Neumann, Laura J.; Star, Susan Leigh; Merkel, Ceceila; Ignacio, Emily and Sandusky, Robert J. (2000). "Digital libraries: Situating use in changing information infrastructure." J. Amer. Soc. for Information Science 51(4): 394-413.
- Deerwester, Scott; Dumais, Susan T.; Furnas, George W.; Landauer, Thomas K.; Harshman, Richard. (1990). "Indexing by Latent Semantic Analysis." J. Amer. Soc. for Information Science 41(6): 391-407.
- Dervin, Brenda and Nilan, Michael S. (1986). "Information Needs and Uses." In: Williams, Martha E. (Ed.). Annual Review of Information Science and Technology. Medford, NJ: Learned Information.
- Frakes, William B. and Baeza-Yates, Ricardo (Eds.). (1992). Information Retrieval Data Structures & Algorithms. Englewood Cliffs, New Jersey: Prentice-Hall.
- Green, Stephan; Marchionini, Gary; Plaisant, Catherine and Shneiderman, Ben. (2000). "Previews and overviews in digital libraries: Designing surrogates to support visual information seeking." J. Amer. Soc. for Information Science 51(4): 380-393.
- Haas, Stephanie W. (1997). "Disciplinary variation in automatic sublanguage term identification." J. Amer. Soc. for Information Science 48(1): 67-79.
- Haas, Stephanie W. (1996). "Natural Language Processing: Toward Large-Scale, Robust Systems. In: Williams, Martha E. (Ed.). Annual Review of Information Science and Technology. Medford, New Jersey: Information Today.
- Ingwersen, Peter. (1992). Information Retrieval Interaction. London: Taylor Graham.
- Kohonen, T; Kaski, S.; Lagus, K.; Salojärvi, J.; Honkela, J.; Paatero, V.; Saarela, A. (1999). "Self organization of a massive text document collection." In Oja, E. and Kaski, S. (Eds.). Kohonen Maps. Amsterdam: Elsevier. pp. 171-182.
- Korfhage, Robert R. (1997). Information Storage and Retrieval. New York: John Wiley & Sons.
- Korfhage, Robert R.; Rasmussen, Edit M.; Belkin, Nicholas; Harman, Donna. (1999). "Invitational Workshop on Information Retrieval Tools." Pittsburgh: School of Information Science. (Available online: <http://www.sis.pitt.edu/%7Eerasmus/workshop.html>).
- Kuhlthau, Carol C. (1993). The Information Search Process. Medford, New Jersey: Ablex.
- Liddy, Elizabeth D. (1997). "Natural Language Processing for Information Retrieval and Knowledge Discovery." Proceedings of the 34th Annual Data Processing Clinic. Urbana,

Illinois: Graduate School of Library and Information Science in the University of Illinois at Urbana-Champaign.

Metcalf, Bob. (1999). "Early Signs Appear of Slowing Internet Growth." InfoWorld, May 10. (Available online: <http://www.infoworld.com/cgi-bin/displayStory.pl?features/990510mids.htm>).

Newby, Gregory B. (1998). "An information access model with a unified approach to data type, retrieval mechanism and information need." Proceedings of the American Society for Information Science Annual Meeting, 475-484. Medford, NJ: Information Today.

PITAC (President's Information Technology Advisory Council). (1999). Information Technology Research: Investing in Our Future. Washington, DC: National Coordinating Office for Computing, Information, and Communications. (Available online: <http://www.ccic.gov/ac/report/>).

Salton, Gerard and McGill, Michael J. (1983). Introduction to Modern Information Retrieval. New York: McGraw Hill.

Schamber, Linda; Eisenberg, Michael; Nilan, Michael. (1991). "A re-examination of relevance: Toward a dynamic, situational definition." Information Processing and Management 26(6): 755-776.

Silverstein, C. and Henzinger, M. (1999). "Analysis of a very large web search engine query log." ACM SIGIR Forum 33(3): 1-8.

vanRijsbergen, C.J. (1979). Information Retrieval. London: Butterworths.

Voorhees, Ellen and Harman, Donna. (2000). "Overview of the Eighth Text REtrieval Conference (TREC-8)." In: Voorhees, Ellen and Harman, Donna (Eds.). The Eighth Text REtrieval Conference (TREC-8). Gaithersburg, Maryland: The National Institute of Science and Technology. (Available online: <http://trec.nist.gov>).

Witten, Ian H.; Moffat, Alistair; Bell, Timothy C. (1999). Managing Gigabytes. San Francisco: Morgan Kaufmann.

Wurman, Richard. (1990). Information Anxiety. New York: Bantam Books.

Biographical Sketch

Gregory B. Newby, Ph.D.

Assistant Professor in the School of Information and Library Science
University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599-3360
919-962-8064, gbnewby@ils.unc.edu

Employment

Assistant Professor, University of North Carolina at Chapel Hill, School of Information and Library Science. (1997-).

Assistant Professor, University of Illinois at Urbana-Champaign, School of Library and Information Science. (1991-1997). Joint appointment as Senior Research Scientist at the National Center for Supercomputing Applications (NCSA).

Education

- Ph.D. in Information Transfer, Syracuse University, 1993. Advisor: Michael S. Nilan. Thesis title: Towards Navigation for Information Retrieval
- M.A. in Communication, State University of New York at Albany, 1988. Advisors: Joseph D. Woelfel and Donald P. Cushman. Thesis title: A Self-Concept Based Approach to Artificial Intelligence, with a Case Study of the Galileo Computer System.
- B.A. in Communication and Psychology, State University of New York at Albany. 1987.

Professional Activities

- Chair of ASIS SIG/ED (Education) 1999-2000. Chair of ASIS SIG/SRT (Storage and Retrieval Technologies), 1992-93; 1994-96. Program Committee, ASIS Annual Meeting 1998.
- Co-Chair, ASIS Mid-Year Meeting 1997.
- Chair, ASIS Awards and Honors Committee, 1993-1995.
- Co-Founder and designer of Prairienet, the Free-Net of East-Central Illinois. 1992- 1997.

Related Publications

“Moving More Quickly Towards Full Term Relations in Information Space.” Gregory B. Newby. 2000. Text REtrieval Conference (TREC-8) Proceedings. Gaithersburg, MD: National Institute of Science and Technology. November 16-19, 1999.

“The strong cognitive stance as a conceptual basis for the role of information in informatics and information system design.” Gregory B. Newby. 1998. Proceedings of the Joint Meeting of the World Multiconference on Systemics, Cybernetics and Informatics (SCI '98) and the 4th International Conference on Informatics Systems Analysis and Synthesis (ISAS '98). Orlando, Florida, July 12-16.

“An information access model with a unified approach to data type, retrieval mechanism and information need.” Gregory B. Newby. 1998. Proceedings of the American Society for Information Science Annual Meeting, 475-484. Medford, NJ: Information Today. Pittsburgh, PA: November 12-16.

“Metric Multidimensional Information Space.” Gregory B. Newby. 1996. Text REtrieval Conference (TREC-5) Proceedings. Gaithersburg, MD: National Institute of Science and Technology. November 20-22.

“Digital Library Models and Prospects. 1996. Proceedings of the American Society for Information Science Mid-Year Meeting. San Diego: May 18-22.

Other Significant Publications

“A Prognosis for Continued Disarray in Electronic Scholarly Communication.” Gregory B. Newby. 1997. The Canadian Journal of Communication. 22: 511-523.

Scholarly Publishing: The Electronic Frontier. 1996. Edited with Robin P. Peek. Cambridge: MIT Press. xxii + 364pp.

“Gesture Recognition Based upon Statistical Similarity.” 1994. Presence 3(3): 236-243.

“The Maturation of Norms for Computer-Mediated Communication.” 1994. Internet Research: Electronic Networking Applications and Policy. 3(4): 30- 38.

Current Collaborators

Debashis Aikat (University of North Carolina), Ann P. Bishop (University of Illinois at Urbana-Champaign), Rolf Doessler (University of Potsdam, Germany), Michael Littman (Duke/AT&T), Gary Marchionini (University of North Carolina), Robin P. Peek (Simmons College).

Advisees

Robert Sumner (University of North Carolina), Rong Tang (SUNY Albany), Michelle Vincow (University of Illinois), Xin Wu (University of Illinois), Kiduk Yang (University of North Carolina).

Advisor

Michael Nilan (Syracuse University).

SUMMARY PROPOSAL BUDGET YEAR 1

ORGANIZATION University of North Carolina at Chapel Hill				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Gregory B Newby				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-mos.		Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR				
1. Gregory B Newby - Dr.	0.00	0.00	1.00	\$ 6,127			
2.							
3.							
4.							
5.							
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0			
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.00	6,127			
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0			
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0			
3. (3) GRADUATE STUDENTS				49,680			
4. (0) UNDERGRADUATE STUDENTS				0			
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0			
6. (0) OTHER				0			
TOTAL SALARIES AND WAGES (A + B)				55,807			
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				3,785			
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				59,592			
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
Server system				\$ 44,000			
TOTAL EQUIPMENT				44,000			
E. TRAVEL							
1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				1,200			
2. FOREIGN				0			
F. PARTICIPANT SUPPORT COSTS							
1. STIPENDS \$ _____				0			
2. TRAVEL _____				0			
3. SUBSISTENCE _____				0			
4. OTHER _____				0			
(0) TOTAL PARTICIPANT COSTS				0			
G. OTHER DIRECT COSTS							
1. MATERIALS AND SUPPLIES				7,060			
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				175			
3. CONSULTANT SERVICES				0			
4. COMPUTER SERVICES				0			
5. SUBAWARDS				0			
6. OTHER				7,125			
TOTAL OTHER DIRECT COSTS				14,360			
H. TOTAL DIRECT COSTS (A THROUGH G)				119,152			
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
UNC indirect (Rate: 45.0000, Base: 68027)							
TOTAL INDIRECT COSTS (F&A)				30,612			
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				149,764			
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.D.7.j.)				0			
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				\$ 149,764			
M. COST SHARING PROPOSED LEVEL \$ 7,500				AGREED LEVEL IF DIFFERENT \$			
PI / PD TYPED NAME & SIGNATURE*			DATE	FOR NSF USE ONLY			
Gregory B Newby				INDIRECT COST RATE VERIFICATION			
ORG. REP. TYPED NAME & SIGNATURE*			DATE	Date Checked	Date Of Rate Sheet	Initials - ORG	

SUMMARY PROPOSAL BUDGET

YEAR 2

ORGANIZATION University of North Carolina at Chapel Hill				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Gregory B Newby				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-mos.		Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR				
1. Gregory B Newby - Dr.	0.00	0.00	1.00	\$ 6,433			
2.							
3.							
4.							
5.							
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0			
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.00	6,433			
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0			
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0			
3. (3) GRADUATE STUDENTS				52,164			
4. (0) UNDERGRADUATE STUDENTS				0			
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0			
6. (0) OTHER				0			
TOTAL SALARIES AND WAGES (A + B)				58,597			
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				3,974			
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				62,571			
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
Server system update and expansion				\$ 38,000			
TOTAL EQUIPMENT				38,000			
E. TRAVEL				2,000			
1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							
2. FOREIGN				0			
F. PARTICIPANT SUPPORT COSTS							
1. STIPENDS \$ _____				0			
2. TRAVEL _____				0			
3. SUBSISTENCE _____				0			
4. OTHER _____				0			
(0) TOTAL PARTICIPANT COSTS				0			
G. OTHER DIRECT COSTS							
1. MATERIALS AND SUPPLIES				6,103			
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				368			
3. CONSULTANT SERVICES				0			
4. COMPUTER SERVICES				0			
5. SUBAWARDS				0			
6. OTHER				7,482			
TOTAL OTHER DIRECT COSTS				13,953			
H. TOTAL DIRECT COSTS (A THROUGH G)				116,524			
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
UNC indirect (Rate: 45.5000, Base: 71042)							
TOTAL INDIRECT COSTS (F&A)				32,324			
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				148,848			
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.D.7.j.)				0			
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				\$ 148,848			
M. COST SHARING PROPOSED LEVEL \$ 7,875				AGREED LEVEL IF DIFFERENT \$			
PI / PD TYPED NAME & SIGNATURE*			DATE	FOR NSF USE ONLY			
Gregory B Newby				INDIRECT COST RATE VERIFICATION			
ORG. REP. TYPED NAME & SIGNATURE*			DATE	Date Checked	Date Of Rate Sheet	Initials - ORG	

SUMMARY PROPOSAL BUDGET YEAR 3

ORGANIZATION University of North Carolina at Chapel Hill				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Gregory B Newby				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-mos.		Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR				
1. Gregory B Newby - Dr.	0.00	0.00	1.00	\$ 6,755			
2.							
3.							
4.							
5.							
6. (0) OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0			
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	1.00	6,755			
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0			
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0			
3. (3) GRADUATE STUDENTS				54,772			
4. (0) UNDERGRADUATE STUDENTS				0			
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0			
6. (0) OTHER				0			
TOTAL SALARIES AND WAGES (A + B)				61,527			
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				4,173			
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				65,700			
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
Server system				\$ 35,000			
TOTAL EQUIPMENT				35,000			
E. TRAVEL				2,100			
1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)							
2. FOREIGN				0			
F. PARTICIPANT SUPPORT COSTS							
1. STIPENDS	\$ 0						
2. TRAVEL	0						
3. SUBSISTENCE	0						
4. OTHER	0						
(0) TOTAL PARTICIPANT COSTS				0			
G. OTHER DIRECT COSTS							
1. MATERIALS AND SUPPLIES				5,348			
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				386			
3. CONSULTANT SERVICES				0			
4. COMPUTER SERVICES				0			
5. SUBAWARDS				0			
6. OTHER				7,856			
TOTAL OTHER DIRECT COSTS				13,590			
H. TOTAL DIRECT COSTS (A THROUGH G)				116,390			
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
UNC indirect (Rate: 45.5000, Base: 73534)							
TOTAL INDIRECT COSTS (F&A)				33,457			
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				149,847			
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.D.7.j.)				0			
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				\$ 149,847			
M. COST SHARING PROPOSED LEVEL \$ 8,269				AGREED LEVEL IF DIFFERENT \$			
PI / PD TYPED NAME & SIGNATURE*			DATE	FOR NSF USE ONLY			
Gregory B Newby				INDIRECT COST RATE VERIFICATION			
ORG. REP. TYPED NAME & SIGNATURE*			DATE	Date Checked	Date Of Rate Sheet	Initials - ORG	

SUMMARY PROPOSAL BUDGET Cumulative

ORGANIZATION University of North Carolina at Chapel Hill				FOR NSF USE ONLY			
				PROPOSAL NO.	DURATION (months)		
PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR Gregory B Newby				AWARD NO.	Proposed	Granted	
A. SENIOR PERSONNEL: PI/PD, Co-PI's, Faculty and Other Senior Associates (List each separately with title, A.7. show number in brackets)				NSF Funded Person-mos.		Funds Requested By proposer	Funds granted by NSF (if different)
	CAL	ACAD	SUMR				
1. Gregory B Newby - Dr.	0.00	0.00	3.00	\$ 19,315			
2.							
3.							
4.							
5.							
6. () OTHERS (LIST INDIVIDUALLY ON BUDGET JUSTIFICATION PAGE)	0.00	0.00	0.00	0			
7. (1) TOTAL SENIOR PERSONNEL (1 - 6)	0.00	0.00	3.00	19,315			
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)							
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0			
2. (0) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	0.00	0.00	0.00	0			
3. (9) GRADUATE STUDENTS				156,616			
4. (0) UNDERGRADUATE STUDENTS				0			
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0			
6. (0) OTHER				0			
TOTAL SALARIES AND WAGES (A + B)				175,931			
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				11,932			
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				187,863			
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING \$5,000.)							
			\$ 117,000				
TOTAL EQUIPMENT				117,000			
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				5,300			
2. FOREIGN				0			
F. PARTICIPANT SUPPORT COSTS							
1. STIPENDS \$ _____			0				
2. TRAVEL _____			0				
3. SUBSISTENCE _____			0				
4. OTHER _____			0				
(0) TOTAL PARTICIPANT COSTS				0			
G. OTHER DIRECT COSTS							
1. MATERIALS AND SUPPLIES				18,511			
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				929			
3. CONSULTANT SERVICES				0			
4. COMPUTER SERVICES				0			
5. SUBAWARDS				0			
6. OTHER				22,463			
TOTAL OTHER DIRECT COSTS				41,903			
H. TOTAL DIRECT COSTS (A THROUGH G)				352,066			
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)							
TOTAL INDIRECT COSTS (F&A)				96,394			
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				448,460			
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.D.7.j.)				0			
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				\$ 448,460			
M. COST SHARING PROPOSED LEVEL \$ 23,644				AGREED LEVEL IF DIFFERENT \$			
PI / PD TYPED NAME & SIGNATURE*			DATE	FOR NSF USE ONLY			
Gregory B Newby				INDIRECT COST RATE VERIFICATION			
ORG. REP. TYPED NAME & SIGNATURE*			DATE	Date Checked	Date Of Rate Sheet	Initials - ORG	

** Personnel and most other costs are increased at 5% per year to reflect anticipated cost of living increases to wages and cost inflation for other items.

** A.1 Senior Personnel – Gregory B. Newby. Dr. Newby is requesting support for one summer month, equivalent to one-ninth of the annual rate.

** B.3 Graduate Students. It is proposed that (3) graduate students be supported for the academic year and 1 summer month during each year of the project. One student will specialize in documentation, two will be primarily programmers; all will have some responsibility for interface design, user support and testing. Cost reflects 20 hours per week for 46 weeks. Additional costs for graduate students at UNC include health insurance (\$832/year for 1999-2000, included in line C) and in-state tuition waiver (\$2262/year for 1999-2000, listed on line G.6).

** C. Fringe benefits are defined as 19% of salary for faculty and health insurance for graduate students (\$832/year in 1999-2000, estimated increase 5%/year).

** D. Equipment. To maintain current equipment over the life of the project, equipment will be replaced or upgraded. UNC considers “equipment” to be specialized tools. This item therefore reflects the main computer server and updates to the server. The server will be used for large-scale testing, and is purposively based on high-end commodity PCs rather than more expensive products from Sun or SGI (note that software will be co-developed for Solaris, IRIX and AIX using servers available elsewhere at UNC). At estimated 2000-2003 prices and performance, plans for this item are:

- a. Fall 2000: Dual- or quad-processor Linux system with 800-1200GB disk space and 1-2GB system memory (processor speed 800-1000Mhz).
- b. Fall 2001: Upgrade processors to contemporary level (estimated 1200Mhz-2000Mhz), expand disk space by 1000-3000GB, expand memory. Depending on availability and cost of equipment at that time, this may be an expansion to the initial server, or an entirely new server.
- c. Fall 2002: Replace server with state-of-the-art, estimated at 4-8 processors (2000+Mhz), 3000-10000GB disk, 4+GB memory. The older server(s) will be kept in service as file and/or compute servers.

** E. Travel. The PI and/or his graduate students will attend one or more of the following conferences to present results: Association for Computing Machinery Special Interest Group in Information Retrieval (ACM SIG/IR), American Society for Information Science (ASIS), and ACM Digital Libraries. Actual attendance depends on time frame and location.

** G.1. Materials and Supplies. UNC defines materials and supplies to include desktop computers and miscellaneous computer equipment, as well as “consumables.”

- a. Consumables. Consumables will be transparencies, pens, hanging file folders, diskettes, software upgrades, etc. The items will be used for presentations, talks etc., in conjunction with this research. \$860 for year 1, \$903 for year 2, \$948 for year 3.

- b. Desktop computers will be used by the GAs for developing, testing, documenting and deploying the IR toolkit. These will run Linux or other Unix variants and be employed as backup Web servers and for development. They will be upgraded or replaced annually to enable performance testing with consumer-grade equipment. \$4000 for year 1, \$3500 for year 2, \$3200 for year 3.
- c. Miscellaneous computer equipment will include rack or shelf storage units, backup media, a network switch, and cabling. Note that the PI's department at UNC will supply the needed network infrastructure and connection to the Internet, systems administration, installation, backups and system security. \$2200 for year 1, \$1700 for year 2, \$1200 for year 3.

** G.2. Publication costs. It is anticipated that books and conference proceedings will be needed.

** G.6. Other costs. This item reflects in-state tuition for graduate student assistants as required by UNC. The 1999-2000 rate is \$2262 annually, with anticipated increases of 5% per year.

** I. Indirect costs. Computed at 45.0% of base for 2000-2001, then 45.5% of base for 2002-2003. The base consists of all budget items except graduate student tuition (line G.6) and equipment (line D).

** M. Cost sharing. The School of Information and Library Science at UNC-CH will perform systems administration, software upgrade and security tasks on all project computers, estimated at .15% of one person's time for the duration of the project (fringe benefits and insurance are included in this estimate, based on an annual salary of \$40,000). 5% per year increases are included.

Current and Pending Support

(See GPG Section II.D.8 for guidance on information to include on this form.)

The following information should be provided for each investigator and other senior personnel. Failure to provide this information may delay consideration of this proposal.	
Investigator: Gregory Newby	Other agencies (including NSF) to which this proposal has been/will be submitted.
Support: <input type="checkbox"/> Current <input checked="" type="checkbox"/> Pending <input type="checkbox"/> Submission Planned in Near Future <input type="checkbox"/> *Transfer of Support Project/Proposal Title: ITR/SW: TeraScale Retrieval	
Source of Support: NSF Total Award Amount: \$ 448,460 Total Award Period Covered: 08/21/00 - 08/20/03 Location of Project: Chapel Hill, NC Person-Months Per Year Committed to the Project. Cal: 0.00 Acad: 0.00 Sumr: 1.00	
Support: <input type="checkbox"/> Current <input type="checkbox"/> Pending <input type="checkbox"/> Submission Planned in Near Future <input type="checkbox"/> *Transfer of Support Project/Proposal Title:	
Source of Support: Total Award Amount: \$ Total Award Period Covered: Location of Project: Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:	
Support: <input type="checkbox"/> Current <input type="checkbox"/> Pending <input type="checkbox"/> Submission Planned in Near Future <input type="checkbox"/> *Transfer of Support Project/Proposal Title:	
Source of Support: Total Award Amount: \$ Total Award Period Covered: Location of Project: Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:	
Support: <input type="checkbox"/> Current <input type="checkbox"/> Pending <input type="checkbox"/> Submission Planned in Near Future <input type="checkbox"/> *Transfer of Support Project/Proposal Title:	
Source of Support: Total Award Amount: \$ Total Award Period Covered: Location of Project: Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:	
Support: <input type="checkbox"/> Current <input type="checkbox"/> Pending <input type="checkbox"/> Submission Planned in Near Future <input type="checkbox"/> *Transfer of Support Project/Proposal Title:	
Source of Support: Total Award Amount: \$ Total Award Period Covered: Location of Project: Person-Months Per Year Committed to the Project. Cal: Acad: Sumr:	

*If this project has previously been funded by another agency, please list and furnish information for immediately preceding funding period.

FACILITIES, EQUIPMENT & OTHER RESOURCES

FACILITIES: Identify the facilities to be used at each performance site listed and, as appropriate, indicate their capacities, pertinent capabilities, relative proximity, and extent of availability to the project. Use "Other" to describe the facilities at any other performance sites listed and at sites for field studies. USE additional pages as necessary.

Laboratory:

Clinical:

Animal:

Computer: At his institution, the PI and his GAs have access to high-end scientific computational facilities. These are not suitable for production work (e.g., building terabyte-scale collections or running Web servers). However, they will be used to insure cross-compatibility of the software toolkit under Linux, Solaris, AIX and IRIX.

Office:

Other: _____

MAJOR EQUIPMENT: List the most important items available for this project and, as appropriate identifying the location and pertinent capabilities of each.

OTHER RESOURCES: Provide any information describing the other resources available for the project. Identify support services such as consultant, secretarial, machine shop, and electronics shop, and the extent to to which they will be available for the project. Include an explanation of any consortium/contractual arrangements with other organizations.

COVER SHEET FOR PROPOSAL TO THE NATIONAL SCIENCE FOUNDATION

PROGRAM ANNOUNCEMENT/SOLICITATION NO./CLOSING DATE (if not in response to a program announcement/solicitation enter NSF 00-2)					FOR NSF USE ONLY	
99-167			02/16/00		NSF PROPOSAL NUMBER	
FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)						
INFORMATION TECHNOLOGY RESEARCH						
DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION	
				003203213		
EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)		SHOW PREVIOUS AWARD NO. IF THIS IS <input type="checkbox"/> A RENEWAL <input type="checkbox"/> AN ACCOMPLISHMENT-BASED RENEWAL		IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES <input type="checkbox"/> NO <input checked="" type="checkbox"/> IF YES, LIST ACRONYMS(S)		
566001393						
NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE			ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE			
University of North Carolina at Chapel Hill			300 Bynum Hall CB 4100 Chapel Hill, NC. 275994100			
AWARDEE ORGANIZATION CODE (IF KNOWN)						
0029744000						
NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE			ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE			
PERFORMING ORGANIZATION CODE (IF KNOWN)						
IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.D.1 For Definitions) <input type="checkbox"/> FOR-PROFIT ORGANIZATION <input type="checkbox"/> SMALL BUSINESS <input type="checkbox"/> MINORITY BUSINESS <input type="checkbox"/> WOMAN-OWNED BUSINESS						
TITLE OF PROPOSED PROJECT ITR/SW: TeraScale Retrieval						
REQUESTED AMOUNT		PROPOSED DURATION (1-60 MONTHS)		REQUESTED STARTING DATE		SHOW RELATED PREPROPOSAL NO., IF APPLICABLE
\$ 448,460		36 months		08/21/00		
CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW						
<input type="checkbox"/> BEGINNING INVESTIGATOR (GPG 1.A.3)			<input type="checkbox"/> VERTEBRATE ANIMALS (GPG II.D.12) IACUC App. Date _____			
<input type="checkbox"/> DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.D.1)			<input type="checkbox"/> HUMAN SUBJECTS (GPG II.D.12)			
<input type="checkbox"/> PROPRIETARY & PRIVILEGED INFORMATION (GPG II.D.10)			Exemption Subsection _____ or IRB App. Date _____			
<input type="checkbox"/> NATIONAL ENVIRONMENTAL POLICY ACT (GPG II.D.10)			<input type="checkbox"/> INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES _____			
<input type="checkbox"/> HISTORIC PLACES (GPG II.D.10)			<input type="checkbox"/> FACILITATION FOR SCIENTISTS/ENGINEERS WITH DISABILITIES (GPG V.G.)			
<input type="checkbox"/> SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.D.12)			<input type="checkbox"/> RESEARCH OPPORTUNITY AWARD (GPG V.H)			
<input type="checkbox"/> GROUP PROPOSAL (GPG II.D.12)						
PI/DP DEPARTMENT SILS			PI/DP POSTAL ADDRESS CB3360 Manning Hall Sch. of Information and Library Science Chapel Hill, NC 275993360 United States			
PI/DP FAX NUMBER 919-962-8071						
NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address		
PI/DP NAME Gregory B Newby	Ph.D.	1993	919-962-8064	gbnewby@ils.unc.edu		
CO-PI/DP						
CO-PI/DP						
CO-PI/DP						
CO-PI/DP						

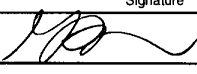
CERTIFICATION PAGE

Certification for Principal Investigators and Co-Principal Investigators:

I certify to the best of my knowledge that:

- (1) the statements herein (excluding scientific hypotheses and scientific opinions) are true and complete, and
- (2) the text and graphics herein as well as any accompanying publications or other documents, unless otherwise indicated, are the original work of the signatories or individuals working under their supervision. I agree to accept responsibility for the scientific conduct of the project and to provide the required progress reports if an award is made as a result of this application.

I understand that the willful provision of false information or concealing a material fact in this proposal or any other communication submitted to NSF is a criminal offense (U.S. Code, Title 18, Section 1001).

Name (Typed)	Signature	Social Security No.*	Date
PI/PI/PI Gregory B Newby		*ON FAST-LANE SUBMISSIONS* SSNs are confidential and are not displayed	2/11/00
Co-PI/PI/PI			
Co-PI/PI/PI			
Co-PI/PI/PI			
Co-PI/PI/PI			

Certification for Authorized Organizational Representative or Individual Applicant:

By signing and submitting this proposal, the individual applicant or the authorized official of the applicant institution is: (1) certifying that statements made herein are true and complete to the best of his/her knowledge; and (2) agreeing to accept the obligation to comply with NSF award terms and conditions if an award is made as a result of this application. Further, the applicant is hereby providing certifications regarding Federal debt status, debarment and suspension, drug-free workplace, and lobbying activities (see below), as set forth in Grant Proposal Guide (GPG), NSF 00-2. Willful provision of false information in this application and its supporting documents or in reports required under an ensuring award is a criminal offense (U. S. Code, Title 18, Section 1001).

In addition, if the applicant institution employs more than fifty persons, the authorized official of the applicant institution is certifying that the institution has implemented a written and enforced conflict of interest policy that is consistent with the provisions of Grant Policy Manual Section 510; that to the best of his/her knowledge, all financial disclosures required by that conflict of interest policy have been made; and that all identified conflicts of interest will have been satisfactorily managed, reduced or eliminated prior to the institution's expenditure of any funds under the award, in accordance with the institution's conflict of interest policy. Conflict which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

Debt and Debarment Certifications

(If answer "yes" to either, please provide explanation.)

Is the organization delinquent on any Federal debt?

Yes

No

Is the organization or its principals presently debarred, suspended, proposed for debarment, declared ineligible, or voluntarily excluded from covered transactions by any Federal department or agency?

Yes

No

Certification Regarding Lobbying

This certification is required for an award of a Federal contract, grant, or cooperative agreement exceeding \$100,000 and for an award of a Federal loan or a commitment providing for the United States to insure or guarantee a loan exceeding \$150,000.

Certification for Contracts, Grants, Loans and Cooperative Agreements

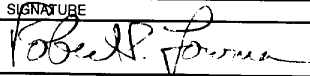
The undersigned certifies, to the best of his or her knowledge and belief, that:

(1) No federal appropriated funds have been paid or will be paid, by or on behalf of the undersigned, to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with the awarding of any federal contract, the making of any Federal grant, the making of any Federal loan, the entering into of any cooperative agreement, and the extension, continuation, renewal, amendment, or modification of any Federal contract, grant, loan, or cooperative agreement.

(2) If any funds other than Federal appropriated funds have been paid or will be paid to any person for influencing or attempting to influence an officer or employee of any agency, a Member of Congress, an officer or employee of Congress, or an employee of a Member of Congress in connection with this Federal contract, grant, loan, or cooperative agreement, the undersigned shall complete and submit Standard Form-LLL, "Disclosure Form to Report Lobbying," in accordance with its instructions.

(3) The undersigned shall require that the language of this certification be included in the award documents for all subawards at all tiers including subcontracts, subgrants, and contracts under grants, loans, and cooperative agreements and that all subrecipients shall certify and disclose accordingly.

This certification is a material representation of fact upon which reliance was placed when this transaction was made or entered into. Submission of this certification is a prerequisite for making or entering into this transaction imposed by section 1352, title 31, U.S. Code. Any person who fails to file the required certification shall be subject to a civil penalty of not less than \$10,000 and not more than \$100,000 for each such failure.

AUTHORIZED ORGANIZATIONAL REPRESENTATIVE		SIGNATURE	DATE
NAME/TITLE (TYPED) Robert P. Lowman, Ph.D., Director			7/14/2000
TELEPHONE NUMBER 919-966-5625	ELECTRONIC MAIL ADDRESS lowman@unc.edu	FAX NUMBER 919-962-6769	

*SUBMISSION OF SOCIAL SECURITY NUMBERS IS VOLUNTARY AND WILL NOT AFFECT THE ORGANIZATION'S ELIGIBILITY FOR AN AWARD. HOWEVER, THEY ARE AN INTEGRAL PART OF THE INFORMATION SYSTEM AND ASSIST IN PROCESSING THE PROPOSAL. SSN SOLICITED UNDER NSF ACT OF 1950, AS AMENDED.