02 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. Submission of this information is voluntary and is not a precondition of award. This information will not be disclosed to external peer reviewers. ***DO NOT INCLUDE THIS FORM WITH ANY OF THE OTHER COPIES OF YOUR PROPOSAL AS THIS MAY COMPROMISE THE CONFIDENTIALITY OF THE INFORMATION.***

PI/PD Name: Thomas Krichel

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 task, 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 will not affect the organization's eligibility for an award. However, information not submitted will seriously undermine the statistical validity, and therefore the usefulness, of information recieved 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 oppurtunities; 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:
Not Listed

REVIEWERS NOT TO INCLUDE:
Not Listed

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 04-23

NSF 04-592 **09/14/04**

FOR CONSIDERATION BY NSF ORGANIZATION UNIT(S) (Indicate the most specific unit known, i.e. program, division, etc.)

IIS - SPECIAL PROJECTS (IIS)

FOR NSF USE ONLY

NSF PROPOSAL NUMBER

0456098

DATE RECEIVED	NUMBER OF COPIES	DIVISION ASSIGNED	FUND CODE	DUNS# (Data Universal Numbering System)	FILE LOCATION
				065933103	

EMPLOYER IDENTIFICATION NUMBER (EIN) OR TAXPAYER IDENTIFICATION NUMBER (TIN)

111633516

SHOW PREVIOUS AWARD NO. IF THIS IS
☐ A RENEWAL
☐ AN ACCOMPLISHMENT-BASED RENEWAL

IS THIS PROPOSAL BEING SUBMITTED TO ANOTHER FEDERAL AGENCY? YES ☐ NO ☒ IF YES, LIST ACRONYM(S)

NAME OF ORGANIZATION TO WHICH AWARD SHOULD BE MADE

Long Island University

AWARDEE ORGANIZATION CODE (IF KNOWN)

0027516000

ADDRESS OF AWARDEE ORGANIZATION, INCLUDING 9 DIGIT ZIP CODE

Long Island University
700 Northern Boulevard
Brookville, NY. 11548

NAME OF PERFORMING ORGANIZATION, IF DIFFERENT FROM ABOVE

PERFORMING ORGANIZATION CODE (IF KNOWN)

ADDRESS OF PERFORMING ORGANIZATION, IF DIFFERENT, INCLUDING 9 DIGIT ZIP CODE

IS AWARDEE ORGANIZATION (Check All That Apply) (See GPG II.C For Definitions)
☐ SMALL BUSINESS ☐ MINORITY BUSINESS
☐ FOR-PROFIT ORGANIZATION ☐ WOMAN-OWNED BUSINESS

☐ IF THIS IS A PRELIMINARY PROPOSAL THEN CHECK HERE

TITLE OF PROPOSED PROJECT **Preserving the informal scholarly record from the Web**

REQUESTED AMOUNT	PROPOSED DURATION (1-60 MONTHS)	REQUESTED STARTING DATE	SHOW RELATED PRELIMINARY PROPOSAL NO. IF APPLICABLE
$ **97,208**	**24** months	**04/02/05**	

CHECK APPROPRIATE BOX(ES) IF THIS PROPOSAL INCLUDES ANY OF THE ITEMS LISTED BELOW

☒ BEGINNING INVESTIGATOR (GPG I.A)
☐ DISCLOSURE OF LOBBYING ACTIVITIES (GPG II.C)
☐ PROPRIETARY & PRIVILEGED INFORMATION (GPG I.B, II.C.1.d)
☐ HISTORIC PLACES (GPG II.C.2.j)
☐ SMALL GRANT FOR EXPLOR. RESEARCH (SGER) (GPG II.D.1)
☐ VERTEBRATE ANIMALS (GPG II.D.5) IACUC App. Date _____
☐ HUMAN SUBJECTS (GPG II.D.6)
Exemption Subsection _____ or IRB App. Date _____
☐ INTERNATIONAL COOPERATIVE ACTIVITIES: COUNTRY/COUNTRIES INVOLVED (GPG II.C.2.g.(iv).(c))
_____
☐ HIGH RESOLUTION GRAPHICS/OTHER GRAPHICS WHERE EXACT COLOR REPRESENTATION IS REQUIRED FOR PROPER INTERPRETATION (GPG I.E.1)

PI/PD DEPARTMENT
Palmer School

PI/PD FAX NUMBER
516-299-4168

PI/PD POSTAL ADDRESS

Brookville, NY 11548
United States

NAMES (TYPED)	High Degree	Yr of Degree	Telephone Number	Electronic Mail Address
PI/PD NAME **Thomas Krichel**	**PhD**	**1999**	**516-299-2527**	**krichel@openlib.org**
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				
CO-PI/PD				

Electronic Signature

CERTIFICATION PAGE

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 debarment and suspension, drug-free workplace, and lobbying activities (see below), as set forth in Grant Proposal Guide (GPG), NSF 04-23. Willful provision of false information in this application and its supporting documents or in reports required under an ensuing 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. Conflicts which cannot be satisfactorily managed, reduced or eliminated must be disclosed to NSF.

Drug Free Work Place Certification

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Drug Free Work Place Certification contained in Appendix C of the Grant Proposal Guide.

Debarment and Suspension Certification

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

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 ☒

By electronically signing the NSF Proposal Cover Sheet, the Authorized Organizational Representative or Individual Applicant is providing the Debarment and Suspension Certification contained in Appendix D of the Grant Proposal Guide.

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 of Lobbying Activities,'' 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 **Kathryn S Rockett**	**Electronic Signature**	**Sep 15 2004 2:48PM**

TELEPHONE NUMBER	ELECTRONIC MAIL ADDRESS	FAX NUMBER
516-299-2523	**kathryn.rockett@liu.edu**	**516-299-3101**

*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.

(1) Intellectual merit

The project will develop and test a new method of preserving scholarly documents from the public Web. Records of author names and title of source documents from standard abstracting and indexing databases are used. After a suitable transformation, these source records are submitted to the Google search engine. Google responses are analysed by purpose-written software. The aim is to find, store and preserve, in an entirely automated way, the full-texts (there may be several versions) of those source documents, as well as other Web pages that relate or refer to the source document. Examples for such other pages include the vita of the author, a reading list mention the document, a mention of it the popular press etc. The stored and preserved pages will enable a completely new breed of scholarly digital libraries. It will provide an enabling force to stimulate informal web publishing in the scholarly community and beyond.

(2) Broader impact

Search engines and the Web continue to revolutionize the access to information by everyone. This project brings this revolution to digital preservation on the one hand and scholarly communication on the other. Scholarly communication on the Web will flourish if there are more reliable ways to access documents. This requires digital preservation. The project provides digital preservation tools. Communities can use these tools to preserve the informally published documents. They can also use the tools to build a comprehensive picture of the impact of their papers as documented through all that has been made available on the public web. At a time when the debate on open access encourages more scholars to use novel ways to publicize their work over the Web, our work will ensure that they are preserved. The informality of the Web is complemented by community-driven digital preservation. The broader impact on society to have more freely available scientific material on the Web can not be overestimated.

TABLE OF CONTENTS

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

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

0. Introduction

The main stumbling blocks to digital preservation are of an economic nature.

A first economic problem relates to the decision to preserve. If we make a costly investment today in preservation activities, we will have the option to keep the preserved asset longer. Every investment project faces an uncertainty problem. But in the case of digital preservation, forecasting ongoing costs and benefits is more difficult than with standard projects such as, say, building a new road. Under these circumstances, it is prudent to reduce costs. This project is concerned with reducing preservation costs.

A second economic problem is that many digital assets worth preserving have been created to generate income for their owners. Asset owners embrace the idea of having the assets preserved but are worried about the idea that this would create an alternative access route that may cut into their revenues. Since preservation without access brings no benefits, attempts at digital preservation are traditionally thwarted by these owners. Under these circumstances, it is prudent to start with digital assets that have primarily been created to advertise something. I am not saying: "let's preserve spam". I am talking, say, about academic papers. Such papers are written without direct profit motives; instead they advertise their authors. Similarly, this funding application has not been written for sale, it advertises my ability to conduct a research project, and it also establishes my claim to having the ideas it contains. In principle, if I make the application available on the Web, I would be pleased to have it preserved *and* be given access to. In recent years, more and more of such "advertising" assets have appeared in public access pages on the Web. Many are very valuable and at risk of disappearing.

In principle, the preservation of the Web is not difficult. One simply needs to take a backup of the Web every so often. Of course, current technology does not allow us to do this at a reasonable cost. This implies that we have to be selective about what we preserve. Yet, how does one determine which Web pages are most valuable and are deserving of preservation?

One approach, pioneered by the Internet Archive at http://www.archive.org/ is to preserve top pages. There are several ways in which top pages can be found. One simple method is to look at pages with addresses that end in a domain name. A more sophisticated alternative is Google's pages rank, see Brin and Page (1997). It is, however, expensive to calculate since it requires information about the link structure of the Web. Whichever method is chosen to decide what top pages are, the results only provide an overall idea of what kind of resources were offered on the early Web and how its services approximately worked. But it does not protect much of the

contents of such services, since the real contents is not on the top level pages. Note that I use the term "page" here, and in the remainder of this document, to describe anything that can be found on the public-access Web. Thus a slide presentation, a PDF document, a picture, etc., can all be pages. But the term page, as used here, implies public access.

This application pioneers a radically different approach. The idea is to start from a set of valuable resources that has been described "off-Web", so to speak. We then look for those valuable resources on the Web, using a search engine for basic searching first, and employing other methods to analyse the search results, second. We want to pioneer this idea in the area of scholarly communication. Here, bibliographies exist that describe scholarly documents. The documents are the valuable resources we want to preserve. We want to preserve them independently of the original place of deposit. And we want to preserve them with their surrounding material. What precisely that surrounding material consists of, is a main issue in this application.

The remainder of application has the following parts. Section 1 has the theory underlying the project. Section 2 briefly discusses some prior work of mine. Section 3 discusses the architecture to be built. Section 4 has elements of a project plan. Section 5 discusses the relevance of the proposal to the funding program. Section 6 has some suggestions for future work based on the proposal. Section 7 discusses the benefits for society. Section 8 concludes.
An appendix has potential objections to the application and my answers.

1. Theory

There is no universally accepted way to refer to a scholarly document. The Serial Item and Contribution Identifier (SICI, see http://sunsite.berkeley.edu/SICI/) is an attempt to create one for the bulk of such documents, the ones published in serials. But it has not been widely deployed. In the meantime, there are "locally" accepted ways to refer to scholarly documents. They are local to a given collection of descriptive data about scholarly documents. Such a collection is usually referred to as an abstracting and indexing database. There are a large number of such databases available. Some of them, such as DBLP (see http:/dblp.uni-trier.de), RePEc (http://repec.org) are run by volunteers and they are freely available. Others, such as ERIC (http://www.eric.ed.gov/) and PubMed Central (http://www.pubmedcentral.nih.gov/) are operated with government subsidies and are also freely available. Finally, others are runs a revenue-generating venture. They usually are not freely available. But they can still be of further use, because there are two basic bibliographic elements that fall under fair use. They are the title of the document and the

name of its authors. These two elements will be referred to, in the remainder of this application, as the pivotal data for a document. No group of authors will choose the same title for more than one scientific document. Thus the pivotal data loosely identifies the document. This simple observation is key to the project.

If we find on the Web a page on which we read some variation of the title (say, one that has a dash to break a word around a line, or that leaves out an accent on a word), and some variation of the author name (say with initials rather than full first names), we can be quite sure that we have found a page related to the document from which we have extracted the pivotal data.

The project aims to develop a system, which, starting from pivotal data about a document, uses search engine-based technology to find on the web all the pages related to it at a given time, classifies such pages both in genre and in time, describes and stores them. There are a whole range of issues that need to be investigated in order to build such a system. First, the source document, for which we have extracted the pivotal data, is really not a just "a" document. Rather, it generates a container. Let's call that container the "pool". At the very simplest, we can think about the pool holding a range of full-text instances of the document. For example, they could be the first draft of the paper or the version as given at a specific conference, etc. As long as the document title remains the same, and the document looks like full-text, we will follow standard practice to consider pages as being full-text instances. Thus, a page on the Web---recall that the pages are anything that we can have access to on the public-access Web---can be a full-text instance in the pool. But "full-text" is only one possible relationship. We will allow other relationships between Web pages and the source document. One key aspect of this application is to develop a set of relationship types, together with software that can reliably detect the relationship type from scanning a page. For the moment, let us call such a relationship type a "bridge". Having worked with a prototype application, here are some example bridges:

"full text"	the page is an instance of the full text of the source document
"display page"	the page contains an abstract of the source document and a link to a full-text on a publisher's web site
"extended abstract"	the page is an extended abstract of the source document
"vita"	the page is a CV of one of the authors of the source document
"pub list"	the page is a list of publications by members of a department or research group, listing the source document

"collection"	the page is a contents list of a journal where the source document appeared or a conference where the source documeent was presented
"reference"	the page is the full-text of another document that contains a reference to the source document
"comment"	the page contains a reference to the document but the page is not the full text of another document

As an example, consider this proposal application. If and when it is available for public access on the Web (I will publish it on my Web site), it belongs to the pool of Brin and Page (1997), with the bridge "comment". If a page on the Web can be found that has a bridge to a source document, it belongs to the pool of the source document. It is a worthy page that needs to be preserved.

The project aim is to implement the construction, maintenance and preservation of pools. The software will search the web for instances of pivotal data as supplied by an abstracting and indexing dataset. It will keep a cache of pages found and assess their worthiness. If a page is found worthy, it is stored and monitored. Under monitoring, the system checks periodically to determine if a new version of the page exists, and stores all new versions, as long as they are still deemed worthy. The result is an ordered set of files and directories that hold pages. This is the data produced by the project. In addition, the project will be making metadata available using the METS scheme. Because of its size and richness of the data and metadata, the project team has no hope to explore all the wealth that will be inside. Instead all the data and metadata will be made available via anonymous ftp, http, OAI and via public rsync, for others to study, or build services on.

2. Results from prior work

I understand that what is proposed here departs radically from previous approaches to digital preservation. Rather than collecting documents locally and storing them, this process does a remote collection of pages that have a relationship to a description of a document that has been supplied externally. The document is mapped to a pool identified by a its handle in the bibliographic database.. Initially, the pool is empty. The project fills the pool with the pages that have a bridge to it.

I would not have much confidence in such a radical innovation myself, had I not seen the opportunities opening up when I built a prototype of the proposed system. This prototype is

unfunded. In it, I examine the DBLP database, see http://www.acm.org/sigmod/dblp/db/, for post 1998 entries. At the time of writing, I only look at papers that have been formally published in conventional academic journals. The conference papers, which form the bulk of DBLP, are to be done later. Compared to the full system proposed in this application, the prototype is fairly limited.

- It only looks for pages that have a full-text bridge.
- It only looks for the full text in pages in Microsoft and Adobe format, not in HTML.
- It only uses paper titles, not author names. It is limited to titles that have five words or more to try to ensure that the right document is being found.
- It takes no account of different versions of the page over time.
- The software is written by me. I am not a trained programmer.

Despite these glaring limitations, I am very encouraged by the findings. I find roughly 25% of the papers are freely available in full text. There is a wealth of auxiliary material to be found for every paper. The accuracy of the results of full-text detection is stunning. I have made some of the data available on Web in a service called DoCIS. This is still a hard-hat area. .But as an example, you can look at:

http://wotan.liu.edu/docis/show?doc=dbl/ijdill/2000_2_4_259_TADLA.html

to see an example of different full-text version of a paper that has been published in a toll-gated academic journal. In fact it is possible to build virtual journals that are composed of early versions of the papers in an academic journal.

3. Architecture

The overall architecture of the system will emphasise a low-cost solution that easily scales with advances in technology. The hardware architecture of the system will be split into two, a front-end and a back-end. The front-end machine downloads pages from the Web. It has to be kept free of IP-based special authentication that are common in the setting of a large organization. Such IP authenticated pages would poison a dataset of public-access resources with proprietary resources. However, it only needs to be a small machine since all it does is download.. The remaining operations will be conducted by the back-end machine. For this purpose I plan to use simple Linux PCs running in parallel. They will be interlinked using Linux Single System Integration technology. They will communicate with the front-end machine using http messages.

The system will use the (by-now) classic LAMP architecture: Linux Apache, MySQL, Perl. It will use the Google API to search the Web. I am satisfied with 1000 queries a day limit for my prototype. In private mail, Google have intimated that I may get more than the 1000 queries per day which is the default limit for the API.

The project software will be written in Perl. Each Perl component will implement a published protocol that documents in detail the workings of the software. At the time of writing, I see requirements for the following protocols:

- front/back protocol: sets out how front-end and back-end machines communicate with each other
- exclusion protocol: sets out how we deal with removal requests from people who do not want to have their material preserved.
- storage protocol: sets out how pages will be stored on the logical disk space.
- search protocol: sets out how searches are conducted by the search engines
- query protocol: sets out how the queries are formed. There will be a limited amount of choices that implementers have to make, and these choices will have to be based on domain-specific knowledge of the implementer. They relate to the decision regarding what titles require searching for and under what circumstance author names have to be brought in to refine a title search.
- response protocol: sets out how the search results are parsed. There will be one principle routine that will be configurable to handle specific results to specific Perl modules for handling.
- scanning protocol: some pages are intermediary pages that lead to further pages that are of potential interest. These will only be cached, not preserved. This protocol will set out how to deal with such pages.
- bridge protocol: will handle the recognition of bridges in pages. It will handle its results to the archival protocol.
- archival protocol: will set out how to store pages when a bridge has been found, i.e. worthy pages. This protocol will only generate metadata about the stored pages. Thus all the metadata requirements will be handled here too.
- monitoring protocol: sets out how worthy pages are monitored.
- export protocol: sets out how the system will be made publicly available.

The project team will publish the protocols. We will try to make all protocols as independent as possible from each other. Separate implementation software will be written for each protocol. Thus, it should be possible to reimplement different parts of the system. How to accomplish this is one of the important research concerns of the project.

4. Project plan

The project will be lead by me, Thomas Krichel, using my research time at Long Island University. During the teachings season, I will hire a researcher to work with me on the system design. The researcher will also do some system administration duties on the Linux cluster. The programming work will all be done by a single programmer based in Novosibirsk, Russia. He has worked for me before. The researcher and I will prepare the protocols and the programmer will implement them. In the Summer, i.e. early May to early September, the researcher will take a break from duties and I will direct work with the programmer in Novosibirsk. I have a Summer home there. This set-up is designed to minimize costs and stretch the time for the project to run for two years. This should allow just enough time to implement the system.

Under this proposal the project will produce a running implementation using the DBLP and RePEc, see http://repec.org datasets. All description of papers will be used. Usage conditions for these datasets explicitly allow for the project to go ahead. During the lifetime of the project and beyond, the project may implement the system for other bibliographic collections if additional resources become available, e.g, if pivotal data and machines are donated.

5. Relationship to "Its' About Time".

I understand that I am not a seasoned member of the digital preservation community. However, the objectives outlined in this proposal address many of the issues presented in the executive summary of the research agenda in Hedstrom (2003). Specifically:

- *Specification, system and tool development, pilot implementation, and evaluation of repository models.*

  This is the heart of what is going to be achieved here. We aim at action research that reveals where the actual problems are.
- *Develop a spectrum of repository architectures.*

There is, of course, s a range of possible architectures. Here are confident that the straight storage of pages will address the vast majority of our requirements. Going beyond that would make the project a lot more expensive.

- *Develop a spectrum of digital archiving services.*
  The project will be a key enabler of other digital archiving services. Essentially, anyone with basis document data can set up an archiving site using our software. But this is not the most important feature of this project. It is important that the result of the archiving activity produces good-end user services. Only if good end-user services on can be built on the archived data, will people have good incentives to engage in the preservation activity. This is key to the application. Services based on archived collection will be interesting both to authors and readers of academic papers. See section 6.
- *Alternative repository models and interoperability*
  This project emphasises a simple, robust approach, In principle, every instance that will run the software written by the project will be self-contained. However, metadata is exported using the OAI protocol for public metadata harvesting. This ensures a basic level of interoperability between initiatives that will take up the technologies that the project will develop.
- *Scalability and cost.*
  These are major concerns of the project. The solutions we propose will be simple enough to ensure that a small organization or even an interested amateur can use them. Small scale activities will have significantly less problems sustaining themselves than large-scale activities that are in constant need of external support. In the long run each project based on our software can migrate quite easily between different disk media since the capacities of these disks is every increasing.
- *Articulating and modeling of curatorial processes.*
  This is where I see much of the innovation in this application. Rather than having digital objects go through a bureaucratic local preservation process at the source, let a thousand flowers bloom on the web and have a combination of tradition librarianship (as evident in the bibliography) and modern search engine technologies do the preservation of what is valuable.
- *Aggregation of items and objects into collections*
  It it important to me that the curatorial process does not merely store data, but also organizes it in a way that aids later information retrieval. For each record in a bibliography the project will

open a container and fill it with material from the web. Therefore, in addition to the traditional organization of the serials literature, a new lower organizational level is created. To link pages on the Web with bibliographic information will considerable enrich the organization of the papes as it is found in their "natural" Web environment.

- *Acquisition and ingest*

  This is an area where the project is unique. Rather than determining the quality of a page at harvesting time, when the machine is out there in very loosely structured Web-land, this application uses the time-honoured services of librarians and lets the computer only do specific jobs. The result is less-likely to be error-prone.

- *Economic and business models*

  I am an economist myself but I don't believe in business models research for something as complex as long-term preservation. I believe that giving people out there ideas (rather than formal models) and open technologies that implement the ideas is likely to be most successful in practice. This project aims to implement a radically new idea that will be the basis of a substantial number of future activities.

6. Further work based on this application

In the past, I have worked a lot people in the digital library. I have been focussing on how large digital libraries, can be built and made freely available by the interaction of many people. Thus I have been concentrated on the people building the library. But of course I need a start-up collection of documents before I can study the people related to them. The RePEc digital library, which I founded in 1997, but which goes back to efforts that I made in 1993, remains my claim that I have shown that very large, unfunded, yet sophisticated digital scholarly digital libraries can be created. Only arXiv.org has a larger discipline-based repository. RePEc realizes that the author is at the center of scholarly communication. Accordingly, at the center of RePEc, lies the RePEc author service. This service allow the authors of the documents to add their personal information as data to the library.

I am very excited about using the type of data that will be pioneered by this project to set up generic author services that allow for two things. First, they allow authors to see the impact of their documents over a long period of time and over a much larger set of media than available in a conventional citation index. A conventional citation index will only contain other scholarly documents. Our dataset will have slides from talks, course descriptions etc. Authors will love to

mine this type of data. And second, while they examine both their documents and the pages that we have found that are related to the documents, we will ask them to improve and correct the our metadata. There is reason to believe that authors will manually check for us every page in the pool for every document that they claim to have written. Thus the dataset will go through a complete human revision, eventually. Thus we will get a very high quality dataset without having to pay for a centralized collection activity.

7. Benefits to society

At the time of writing, there has been much discussion about open access to scholarly documents. There is an emerging infrastructure of open-access scholarly journals that is becoming available. In addition, universities have been trying, with limited success, to set up institutional archives. In the meantime, there is a flourishing culture of academic self-publishing on the Web. But it has no technical backup and is not suitable for building advanced services such as peer review, because the documents are in the hands of the authors and may get changed. Digitally preserved copies not only serve as technical backup, they also facilitate alternative peer review. This in turn will stimulate the supply of open access scholarly documents and make the Web, backed up by our preservation system, a public-access library for most scholarly documents. This will be an unprecedented public good to achieve.

8. Conclusions

This application describes a simple project. We make pages on the Web more long-lived. A human determines what kind of objects are valuable. An engine finds candidate pages. Our project analysis them, stores and monitors them. Apart from my prototype, there is no project that implements this simple set-up. As much as search engines have transformed information retrieval on the Web, so could they transform long-term archiving.

Appendix: Potential objections and my responses

1) *Would not be easier to download and store papers from academic departments and store them?*

In the US, academic pages can be quite easy found by looking at domains that end in .edu. Unfortunately, this argument does not generalize across the world. Many countries do not have such divisions in the domains. Even in the US, a lot of personal pages are maintained in .edu,

and a lot of low-quality data is to be found on home pages of students found in this domain. Even if a .edu download and repeated store were technically and economically feasible, it would not add any organizational structured to the contents. By contrast, this project adds intelligence to the scholarly Web that is totally unique.

2) *There are not may full-text papers available outside a hand-full of areas.*
Yes, the perceived wisdom is that there is not much out there apart from the preprints and working papers disciplines, i.e. Physics, Mathematics, Computer Science, and Economics. But this perceived wisdom has not been empirically tested. This application tests it. And if we don't find many worthy documents now, this does by no means ensure that we won't find them in the future. As the whole of scholarly communication is moving towards open access, more and more material is likely to appear. Open access is making a system like our more valuable day by day,

3) *Isn't this like citation analysis? CiteSeer is already doing this.*
Not as far as I am aware. Their system extracts citations from scientific papers. If the citation has a URL, it fetches the cited paper, indexes it etc. This is obviously not the same as what I plan here, because my work is based on a bibliographies. But I would like to note that of course, the resulting full-text gather with our project could be used in further projects such as CiteSeer. Thus our work is an ideal complement to CiteSeer.
4) *You have not addressed the long-run preservation in your project*
In general, it does seem difficult to develop long-run preservation activities from short-lived projects. I feel reluctant to make promises that I am not sure about fulfilling. Just consider the lifetime of past funded digital library activities. A lot do not stretch much further than the day the funding was used up, despite of what what they claim in funding application documents. I don't want this project to end any time soon. Therefore, my key idea is an enabling technology that reduces costs. It will allow others a quick start to digital preservation activities. I expect that there will be a lot of independent digital preservation activities based on our software. I also expect that for each implementation, the amount of material to be preserved is quite small, in the order of a few Terabytes. Therefore, we can hope that the datasets will be safely migrated. In the same way I have maintained RePEc for over ten years, by finding a growing number of volunteers while at the same time reducing my own input. I hope to make a contribution to digital

preservation with a new method, leaving it to others to later maintain systems that implement my idea.

5) *What about copyright?*

The project's work implies no copyright transfer. Folks who don't like their work to the in the preserved collection can either use the robots.txt exclusion protocol, which we will honour, or contact us so we will remove their pages and prevent them from coming in. From discussions that I have had with the Internet archive, that is the way they work. It seems to be working well.

REFERENCES CITED

Brin, Sergey, and Lawrence Page (1997), "The Anatomy of a Large-Scale Hyper-textual Web Search Engine" available at http://www-db.stanford.edu/~backrub/google.html

Hedstrom, Margaret (2003) "Its' about time: Research Challenges in Digital Archiving and Long-term Preservation", final report, NFS/LoC Workshop on Research Challenges in Digital Archiving and Long-Term Preservation held April 12-13, 2002.

Thomas Krichel http://openlib.org/home/krichel

Education

1999 PhD in Economics, University of Surrey 1990 MA, University of Exeter

1989 Magistere d'Economie, Paris I, ENS (Ulm) & EHESS 1996 D.E.U.G., Toulouse I

Academic Positions

since 2001 Assistant professor at the Palmer School of Library and Information Science, Long Island University, New York

2000 Visiting professor at Hitotsubashi University, Tokyo

1993-2001 Lecturer in Economics at the University of Surrey

1992-1993 Houblon-Norman research assistant at the Bank of England and Keele University

1990 -1992 Building Societies Trust Research Assistant at Loughborough University

Publications

(i) Most closely related to the proposal:

This proposal is a totally new approach to digital libraries. Although I have been working with the prototype for about one year, I have not published a formal paper about it at this time. Thus I have decided to list nothing here.

(ii) Others:

Heting Chu and Thomas Krichel (2003), "NEP: Current Awareness Service of the RePEc Digital Library", D-Lib Magazine, December 2003, vol. 9, no. 12, http://www.dlib.org/dlib/december03/ chu/12chu.html

Alison Buckholtz, Raf Dekeyser, Melissa Hagemann, Thomas Krichel, and Herbert Van de Sompel (2003), "Open access: Restoring scientific communication to its rightful owners", European Science Foundation Policy Briefing number 21, http://www.esf.org/publication/157/ESPB21.pdf

Thomas Krichel and Simeon M. Warner (2001), "Design of a metadata framework to support scholarly communication", presented at the International Conference on Dublin Core and Metadata Applications in Tokyo, Japan, October 24 to 26, http://www.nii.ac.jp/dc2001/proceedings/ product/paper-22.pdf

Jose Manuel Barrueco Cruz, Markus J.R. Klink and Thomas Krichel (2000) "Personal data in a large digital library", presented at the Fourth European Conference on Research and Advanced Technology for Digital Libraries in Lisbon, September 18 to 21, http://openlib.org/home/krichel/papers/phoenix.a4.pdf

Herbert Van de Sompel, Thomas Krichel, Michael L. Nelson, Patrick Hochstenbach, Victor M. Lyapunov, Kurt Maly, Mohammad Zubair, Mohamed Kholief, Xiaoming Liu and Heath O Connell (2000), The UPS Prototype: An Experimental End-User Service across E-Print Archives , D-lib Magazine, vol. 6, no. 2, http://www.dlib.org/dlib/february00/vandesompelups/02vandesompel-ups.html

Synergistic activities

Founder and principal coordinator of RePEc, the largest decentralized academic digital library in the world

Co-founder of the Open Archives Initiative, which implements the principals behind RePEc on a more general level

Founder of rclis, a project that imitates RePEc for the computing and library and information science fields (in building-up phase)

Collaborators & Other Affiliations

(i) Collaborators: Nisa Bakkalbasi (SUNY Purchase), Jose Manuel Barrueco Cruz (University of Valencia), Alison Buckholtz (SPARC), Heting Chu (Long Island University), Raf Dekeyser (University of Leuven), Melissa Hagemann (Open Society Institute), Patrick Hostenbach (LANL), Mohamed Kholief (Old Dominion University), Markus J.R. Klink (oose.de), Michael E.D. Koenig (Long Island University), Paul Levine (University of Surrey), Xiaoming Liu (Old Dominion University), Victor M. Lyapunov (Siberian Branch of the Russian Academy of Sciences), Kurt Maly (Old Dominion University), Michael L. Nelson (Old Dominion University), Heath O'Connell (Fermilab), Sergey I. Parinov (Siberian Branch of the Russian Academy of Sciences), Jeremiah C. Trinidad (Columbia University), Herbert Van de Sompel (LANL), Simeon M. Warner (Cornell University), Mohammad Zubair (Old Dominion University)

(ii) PhD advisor: Paul L. Levine (University of Surrey)

(iii) PhD advisees: Marco Catenaro (European Central Bank)

SUMMARY PROPOSAL BUDGET YEAR 1

FOR NSF USE ONLY		
PROPOSAL NO.	DURATION (months) Proposed	Granted
AWARD NO.		

ORGANIZATION: **Long Island University**

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR: **Thomas Krichel**

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-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Thomas Krichel - assistant professor**	0.00	0.00	0.00	$ 0	$
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	0.00	0	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0	
2. (1) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	9.00	0.00	0.00	18,000	
3. (0) GRADUATE STUDENTS				0	
4. (0) UNDERGRADUATE STUDENTS				0	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				18,000	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				6,102	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				24,102	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
Linux cluster with 5 identical computers $ 7,000					
TOTAL EQUIPMENT				7,000	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				0	
2. FOREIGN				1,000	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				0	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				12,000	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				300	
TOTAL OTHER DIRECT COSTS				12,300	
H. TOTAL DIRECT COSTS (A THROUGH G)				44,402	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
salaries & wages (Rate: 24.7000, Base: 18000)					
TOTAL INDIRECT COSTS (F&A)				4,446	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				48,848	
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.C.6.j.)				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				$ 48,848	$

M. COST SHARING PROPOSED LEVEL $ 0 AGREED LEVEL IF DIFFERENT $

PI/PD NAME: **Thomas Krichel**

ORG. REP. NAME*: **Kathryn rockett**

FOR NSF USE ONLY		
INDIRECT COST RATE VERIFICATION		
Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET YEAR 2

FOR NSF USE ONLY		
PROPOSAL NO.	DURATION (months) Proposed	Granted
AWARD NO.		

ORGANIZATION: **Long Island University**

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR: **Thomas Krichel**

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-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Thomas Krichel - assistant professor**	0.00	0.00	0.00	$ 0	$
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	0.00	0	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0	
2. (1) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	9.00	0.00	0.00	20,000	
3. (0) GRADUATE STUDENTS				0	
4. (0) UNDERGRADUATE STUDENTS				0	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				20,000	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				6,420	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				26,420	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.)					
TOTAL EQUIPMENT				0	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				2,000	
2. FOREIGN				0	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				1,000	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				14,000	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				0	
TOTAL OTHER DIRECT COSTS				15,000	
H. TOTAL DIRECT COSTS (A THROUGH G)				43,420	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE) **salaries & wages (Rate: 24.7000, Base: 20000)**					
TOTAL INDIRECT COSTS (F&A)				4,940	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				48,360	
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.C.6.j.)				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				$ 48,360	$

M. COST SHARING PROPOSED LEVEL $ 0 AGREED LEVEL IF DIFFERENT $

PI/PD NAME: **Thomas Krichel**

ORG. REP. NAME*: **Kathryn rockett**

FOR NSF USE ONLY		
INDIRECT COST RATE VERIFICATION		
Date Checked	Date Of Rate Sheet	Initials - ORG

 ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

SUMMARY PROPOSAL BUDGET

Cumulative

FOR NSF USE ONLY		
PROPOSAL NO.	DURATION (months) Proposed	Granted
AWARD NO.		

ORGANIZATION: **Long Island University**

PRINCIPAL INVESTIGATOR / PROJECT DIRECTOR: **Thomas Krichel**

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-months CAL	ACAD	SUMR	Funds Requested By proposer	Funds granted by NSF (if different)
1. **Thomas Krichel - assistant professor**	0.00	0.00	0.00	$ 0	$
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	0.00	0	
B. OTHER PERSONNEL (SHOW NUMBERS IN BRACKETS)					
1. (0) POST DOCTORAL ASSOCIATES	0.00	0.00	0.00	0	
2. (2) OTHER PROFESSIONALS (TECHNICIAN, PROGRAMMER, ETC.)	18.00	0.00	0.00	38,000	
3. (0) GRADUATE STUDENTS				0	
4. (0) UNDERGRADUATE STUDENTS				0	
5. (0) SECRETARIAL - CLERICAL (IF CHARGED DIRECTLY)				0	
6. (0) OTHER				0	
TOTAL SALARIES AND WAGES (A + B)				38,000	
C. FRINGE BENEFITS (IF CHARGED AS DIRECT COSTS)				12,522	
TOTAL SALARIES, WAGES AND FRINGE BENEFITS (A + B + C)				50,522	
D. EQUIPMENT (LIST ITEM AND DOLLAR AMOUNT FOR EACH ITEM EXCEEDING $5,000.) $ 7,000					
TOTAL EQUIPMENT				7,000	
E. TRAVEL 1. DOMESTIC (INCL. CANADA, MEXICO AND U.S. POSSESSIONS)				2,000	
2. FOREIGN				1,000	
F. PARTICIPANT SUPPORT COSTS					
1. STIPENDS $ 0					
2. TRAVEL 0					
3. SUBSISTENCE 0					
4. OTHER 0					
TOTAL NUMBER OF PARTICIPANTS (0) TOTAL PARTICIPANT COSTS				0	
G. OTHER DIRECT COSTS					
1. MATERIALS AND SUPPLIES				1,000	
2. PUBLICATION COSTS/DOCUMENTATION/DISSEMINATION				0	
3. CONSULTANT SERVICES				26,000	
4. COMPUTER SERVICES				0	
5. SUBAWARDS				0	
6. OTHER				300	
TOTAL OTHER DIRECT COSTS				27,300	
H. TOTAL DIRECT COSTS (A THROUGH G)				87,822	
I. INDIRECT COSTS (F&A)(SPECIFY RATE AND BASE)					
TOTAL INDIRECT COSTS (F&A)				9,386	
J. TOTAL DIRECT AND INDIRECT COSTS (H + I)				97,208	
K. RESIDUAL FUNDS (IF FOR FURTHER SUPPORT OF CURRENT PROJECTS SEE GPG II.C.6.j.)				0	
L. AMOUNT OF THIS REQUEST (J) OR (J MINUS K)				$ 97,208	$

M. COST SHARING PROPOSED LEVEL $ **0** AGREED LEVEL IF DIFFERENT $

PI/PD NAME: **Thomas Krichel**

ORG. REP. NAME*: **Kathryn rockett**

FOR NSF USE ONLY		
INDIRECT COST RATE VERIFICATION		
Date Checked	Date Of Rate Sheet	Initials - ORG

C ***ELECTRONIC SIGNATURES REQUIRED FOR REVISED BUDGET**

Budget justification

Bill Arms has told me that I have reputation to get a lot done with little resources. I intend to keep this reputation. Ingenious, though sometimes unusual organization is key to the way that I work.

At the present time, Long Island University do not offer professors working on research projects the option to use research funds to reduce their teaching load. Thus I will have to lead the project out of my general research time. That is not necessarily bad news. My role will be the overall system design. For that spontaneous ideas are more important than continuous nine-to-five labouring.

I plan to hire a team of two people. I have been working together with both of them before and I trust that their talents complement mine. The first is Angela Cornwell. She is one of our recent MLS graduates. She has a computer science background. She will be working as a researcher. She will be based at the CW Post Campus of Long Island University. Her income is counted as wages in the budget. The second is Roman Shapiro. He is about to graduate from the Information Science program at Novosibirsk State University. He will be working as a programmer. He will be based in Novosibirsk. His income is counted as consultancy in the budget.

In order for Angela and me to communicate with Roman, I have budgeted some expenses for phone calls. Using pre-paid cards, one can talk to Novosibirsk from NYC for $1 every 20 minutes. $300 should give us 100 hours of phone time which seems amply sufficient.

For the back-end machine, I need to buy a large computer but have little funds to buy it. Thus, I will use a cluster of Linux machines as the back-end machine. OpenSSI will be used to build a single system image cluster. Every machine will have a four 250 Gigabyte IDE hard drives, making for a total of 5 Terabytes of disk space. Some of the 20 IDE disks may fail. That is why I budget $1000 in the second year on spare parts. For the front-end machine, I will use a computer at my home. It is already sitting there, so it has not been counted in the budget.

In the first year, an additional item $1000 appears for foreign travel. This is the budget item to attend the PI meeting in Summer 2005. Recall that in the Summer I will be living in my secondary residence in Novosibirsk and work with Roman directly. I do not charge the project for my travel to Novosibirsk and back. But I will use project funds to

get from Novosibirsk to Washington DC and back.

In the second year a $2000 item appears for domestic travel. This will be used towards the project end and maybe after it to promote the results and seek partners for further ventures based on the system.

That's all!

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: **Thomas Krichel**

Other agencies (including NSF) to which this proposal has been/will be submitted.

Support: ☒ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *Transfer of Support

Project/Proposal Title: **Academic Contributor Information System**

Source of Support: **Open Society Institute**

Total Award Amount: $ **50,000** Total Award Period Covered: **03/01/03 - 03/01/05**

Location of Project: **Long Island University**

Person-Months Per Year Committed to the Project. Cal: **1.00** Acad: **0.00** Sumr: **0.00**

Support: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *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: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *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: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *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: ☐ Current ☐ Pending ☐ Submission Planned in Near Future ☐ *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: Summ:

*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: **We will buy 5 computers and cluster them using OpenSSI. The cluster will be made available on the Internet at the University Computer Center.**

Office:

Other: **The project will use the university network, as well as the Internet connection of the principal investigator.**

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 which they will be available for the project. Include an explanation of any consortium/contractual arrangements with other organizations.