One uncertainty in the overall project which makes precise definition of costs difficult at this stage is the lack of any specifics about the function and structure of the image servers to be installed at each institution. Thus, costs in this area must be regarded as highly preliminary.

Estimated costs:

Scanner: $12,000

Staff for Scanning: $10,000/year (at least first year)

Staff for Database Storage Investigation: $60,000 (one year) (includes research assistants plus primary investigators)

Staff for initial exploration of multimedia/complex object transfer (one year): $30,000

Staff for image server architecture/protocol work: $20,000 (one year).

The creation of the bibliographic records and their subsequent distribution and incorporation into public access databases is viewed primarily as a production effort, with little research component. Estimated costs for this activity are as follows:

Creation of records: $5,000/year
Development of mapping programs: $10,000 (one–time)
Initial setup of public–access database: $20,000 (one–time)

Total costs (excluding overhead) for bibliographic component: $35,000.

UC would probably absorb the operating costs of the public access database, as long as it did not get too large and use remained at reasonable levels. If very large numbers of retrospective records were incorporated into the database at some point, some modest contribution towards disk space to house the database might be needed.

One point worth noting is that access to this database would be not only through a terminal–oriented interface (accessible through TELNET) but also through the ANSI/NISO Z39.50 computer–to–computer protocol for information retrieval, which would allow workstation clients to access the database readily through locally–developed user interfaces (such as the one under development as part of the Carnegie Mellon Mercury Project).

Access to Report Contents

In contrast to the creation and distribution of the bibliographic database, creating and providing access to report content involves a number of important research aspects.

The UC Berkeley Computer Science Department already has an optical disk jukebox providing very large archival storage as part of the Mammoth Project; the Postgres Database system developed at UC Berkeley has already been interfaced to the optical storage subsystem. Consequently, except for a scanner, the majority of the hardware base is already in place.

Professor Michael Stonebraker at the UC Berkeley Computer Science department and Dr. Clifford Lynch at the Division of Library Automation are particularly interested in exploring the use of the Postgres system as a means of storing and providing access to source documents, either in image or (preferably) textual formats. Dr. Lynch is also interested in protocols and architectures for image servers. There is also considerable interest in extending this project to allow the transfer of complex electronic documents containing elements such as datasets or executable programs, although work in this area has not yet been defined in any detail; perhaps, for the present, the best way to look at the project is as an enabling framework that could support research in such areas, which are of great long–term importance. Finally, it seems likely that expertise from the UC Berkeley School of Library and Information Science could also be brought into the project, although discussions on this have not yet taken place.

records used almost universally by the library community to describe material of this type. Part of the project plan calls for encouraging as many other institutions as possible to contribute such bibliographic records for their technical reports once the system is working at the original consortium members.

Initially, electronic mail (probably using a mail reflector) has been proposed as a distribution mechanism; longer–term, as the project scales up, it is likely that this distribution method will require re–examination.

The second component of the project calls for the members of the consortium to capture the actual content of the technical reports in electronic form, with bit–mapped images being used as a lowest common denominator, and with additional formats such as ASCII or marked–up text being optional. For bibliographic records generated by the consortium members, links would be created that allowed retrieval of report contents by other members of the consortium. Each participating consortium member would store its own reports on a local server. Still largely undefined is how one accesses these servers (e.g. is the server an X application that communicates with end–user workstations, an FTP–based server, or a server running some as–yet–undefined image retrieval protocol?)

UC Participation Plan

The bibliographic component.

UC Berkeley would prepare records in the text–based format described for its computer science technical reports; if the current format (or a near variant) is adopted by the participants, it should be feasible to do this with an off–the–shelf text editor.

When a sufficiently large critical mass of bibliographic records is available from the consortium participants and other institutions that may begin to generate such records, the Division of Library Automation would attempt to make them available to the overall Internet community through its MELVYL online catalog.

In addition, the Division of Library Automation would develop programs to map between these text–formatted bibliographic records and the MARC format, and would make these programs available to other interested parties. If there was sufficient demand, the Division of Library Automation might also consider the development of a network–accessible record remapping service to convert from one format to another.

There is as yet no consensus within the project as to the need for a central record validation and redistribution point; however, should the need for such a function arise, the Division of Library Automation has extensive experience in managing multiple autonomous input and output record streams against a central reference database and would be well–suited to explore undertaking such a role.

An Informal Proposal for University of California Participation in the Network–based Distribution of Computer Science Technical Reports

Clifford A. Lynch
June 15, 1990

DRAFT

Disclaimer: This is an informal proposal indicating UC's interests in this project and estimates of costs and timelines, as well as some discussion of how work at UC might interrelate to that proposed by other potential participants in the consortium. This proposal has NOT been processed through the University of California's Contracts and Grants office. Costs in this proposal do NOT include University of California overhead costs, for example.

OVERVIEW

A small consortium of research universities, including Carnegie Mellon, Stanford, University of California, and MIT, is proposing a project to improve access to the technical report literature in computer science, with emphasis on reports generated by the participant institutions. This project will provide valuable experience with an operational prototype of several technologies necessary to construct electronic libraries.

Within the University of California, the participants in the project would be the Division of Library Automation, a unit within the Office of the President's Information Systems and Administrative Services which has UC system–wide responsibility for public access bibliographic systems and multi–campus library automation, and the University of California at Berkeley's computer science department. At later stages of the project it is hoped that computer science departments at other University of California campuses might become involved; however, no discussions on this topic have as yet taken place with other campuses, the feeling being that it will be easier to get the project up and running without the additional management complexities of multi–campus participation.

The project, as currently defined, has two basic components. Each consortium institution will generate bibliographic records describing its technical reports and make these available for distribution over the Internet to other institutions. The precise format of these records has not yet been defined, but there seems to be some consensus emerging that these should be simple, textually based structured records, and a draft format proposal has been developed. For a number of applications, it will be useful to be able to map data between this format and the MARC (Machine–Readable Cataloging)

Scanning

Stanford would like to acquire scanning equipment for this project. Since reports are produced all year around it makes sense to make them available immediately and not batch them to a service bureau (about $10,000).

Two Developer workstations, similar to the above but with more disk, etc. (about $7,500/ea = $15,000).

One printer station, LPS–20, which prints 20 pages per minute and prints on both sides (about $17,000).

Servers

Stanford production of reports (which includes some theses) in computer science is estimated as:

	number	average pages	total pages
Computer Science Dept.	100	56	5600
Knowledge–Systems Lab	89	56	4984
Computer Systems Lab	88	56	4928
CSLI (languages & info)	19	56	1064
TOTAL (pages per year)			16,576

Storage requirement in G4 Fax format:

16576 pages at 50,000 bytes per page = 829 Mbytes/year

Allowing double capacity for indexes, experimental databases, errors in data, etc., the storage requirement for five years of reports are:

Stanford: 6–7 Gbytes

The server requirement with disk storage (for five years) would be a DecStation 5000 with 10 Gbyte disk storage. This would mean attaching fifteen 665MB disks at a cost of $58,4380. In addition, one 665MB disk for bibliographic and full–text storage is needed for an additional $3,892. When the disk cost is added to the DecStation 5000 the total is over $77,049.

The more desirable configuration would be for Stanford to use an optical server, such as the Epoch 32A2 server with 30 Gbyte WORM storage. (about $109,480; Stanford discount price). Epoch's smallest optical server is 30 MB with two 760 MB magnetic disk drives and 2 WORM drives which allow serve and back–up simultaneously. In addition, a DecStation 5000cx with 16MB of memory, 997MB of disk, 19 inch monochrome display would be needed. (about $18,669). TOTAL cost $128,149. Alternatively, another optical server could be chosen that is NFS–based and would provide access to the Newton retrieval servers via NFS.

COST ESTIMATES FOR STAFF AND EQUIPMENT*

Staff Requirements

1. Collection of material, including source documents. (.25 FTE)

   publications coordinator ($9,000)

2. Software implementation, integration and support for retrieval, display, of full–text modules (1 FTE)

   programmer/analyst ($80,000)

3. Integration of image and library bibliographic modules (.5 FTE)

   programmer/analyst ($40,000)

4. Management of the project. (.25 FTE)

   manager ($20,000)

TOTAL ($149,000)

Hardware Requirements

No.	Description	Unit cost	Total cost
2	Image capture workstations	$5,000	$10,000
2	Indexing/image delivery software development workstations	7,500	15,000
1	High resolution, high speed document image printer	17,000	17,000
1	Image server	128,000	128,000
1	Scanner	10,000	10,000
1	Fiber optic network node	20,000	20,000
	TOTAL		200,000

ASSUMPTIONS USED TO GENERATE COST ESTIMATES

Workstations

Two DecStation 3100, with 19 inch monochrome display, 16 Mbyte memory, 100 Mbyte disk, Ethernet (about $5,000/ea = $10,000). One for the libraries and one for the CS Dept.

---

* Indirect cost not included

Stanford University

Computer Science Technical Report Project Proposal

Executive Summary

May 1, 1990

LONG–TERM GOAL

A sizable collection of computer science technical reports capable of being searched and displayed in their entirety and in real time across a national telecommunications network.

SCHEDULE, MILESTONES, TASKS

May 1990 through November 1990 — 6 months

Creation and exchange of bibliographic data

- Set up electronic–mail based system for sharing announcements of new technical reports.

May 1990 through May 1991 — 12 months

Capture, index, display full–page images (G4, PostScript)

- storage on image servers
- link to bibliographic index
- real–time retrieval across the NSFNet

November 1991

Timely capture and storage of all suitable documents

- use of retrieval/display software by researchers

Future

Consolidate and expand

- turn prototype project into continuing system
- extend the system to meet medium term goals

Laboratory for Computer Science and Artificial Intelligence Laboratory	
2 access workstations for the Computer Science reading room	$20,000
1 input workstation for LCS publications	10,000
1 input workstation for AI publications	10,000
Image scanner	21,000
Image storage server array (8 Gbytes, magnetic)	119,000
Image backup/archive server (64 Gbytes, optical)	42,000
Catalog server (1 Gbyte, magnetic)	50,000
1 workstation for programming staff & students	12,000
TOTAL	356,000

Additional hardware for the second level:

Laboratory for Computer Science	
3 workstations for students and staff	$36,000
experimental index server	50,000
TOTAL	86,000

V. BUDGET

12 months	First level	Second level	Total
Hardware	$356,000	$86,000	$442,000
Staff	167,000	145,000	312,000
Travel, M&S	33,000	15,000	48,000
Overhead	116,000	64,000	180,000
TOTAL	672,000	310,000	982,000

For an 18–month budget, all categories except Hardware increase by 50%.

VI. SUPERVISION

The investigation under this proposal will be carried out under the supervision of,

For the MIT Library System:
T. Gregory Anderson
Associate Director for Systems and Planning

For the MIT Laboratory for Computer Science:
Jerome H. Saltzer
Professor of Computer Science

would be carried out by graduate and undergraduate students of the Laboratory for Computer Science.

III. STAFF

For the first level (Percentages are of full–time equivalents):

Library System

80%	technical person, expert on scanning equipment
20%	supervision

Laboratory for Computer Science

10%	bibliographic assistant in publications office
100%	programmer and computer system wizard
40%	two undergraduate student programmers
20%	coordinator at reading room
30%	person to maintain the shared bibliographic file
20%	supervision
10%	support

Artificial Intelligence Laboratory

10%	bibliographic assistant in publications office

Additional staff for the second level:

Library System

(none proposed)

Laboratory for Computer Science

200%	2 graduate research assistants
80%	4 undergraduate programmers
30%	supervision
15%	support

IV. HARDWARE

For the first level (based on round numbers of 200 publications/year, 12000 pages scanned, for four years, plus an equal size backlog):

Library System

Upgrade of existing fiche camera to permit scanning	$20,000
Workstation for review of scanned input	10,000
2 access workstations for Engineering and Science libraries	20,000
2 laser printers for Engineering and Science libraries	10,000
Workstation for development	12,000

II.2.A. Organizationally, this project represents an opportunity for MIT to try out integration of its several information production and dissemination activities in a way that has been frequently discussed but never before feasible, because of differing goals, standards, and levels of automation. Currently, MIT technical reports and theses are handled by many different ad hoc mechanisms, most of which involve being on–line at one stage or another, but none of which are completely electronic. The various laboratories and departments produce them, the MIT Library System catalogs, archives, and in some cases distributes these reports, and special offices such as Industrial Liaison bring them to the attention of their member communities. Students discover them, but often only accidentally and too late to be of value.

The medium of an electronic library with on–line images of the pages of the technical reports is a vehicle that can bring these diverse production and dissemination activities into an integrated and, we hope, much more effective form. By building this system, MIT expects to be able to step well ahead of where it would be by natural evolution of the current diversified approach. The system will serve as a foundation for future development and expansion of information services.

MIT would extend the basic consortium proposal, which is restricted to technical reports published after the project begins, by including all library–deposited theses in computer science (this is a modest extension, because a majority of such theses are already issued as technical reports), and by gradually incorporating the entire 25–year history of MIT Computer Science technical reports. If time and resources permit, MIT would add also two additional series that are widely looked for but hard to locate, L.C.S. Technical Memos and A. I. Laboratory Memos, as well as a small number of monographs that have status equivalent to laboratory technical reports.

II.2.B. Technically, there will be a focus on two areas: the special problems of managing large scale archival storage, and on discovering simple, low–effort, high–payoff ideas for making immediate widespread use of the images and associated text.

For this purpose MIT will collect not only page images but also whatever machine–readable text also exists for the technical reports, with the intent of trying to convert the many existing forms into a common form suitable for experimenting with full–text search. MIT will also retrospectively insert three items for all older MIT technical reports: a catalog record, an ACM subject category record, and a machine–readable abstract. These materials all exist in machine–readable form today, but in different systems and for different purposes.

At the present time, both levels would be carried out as a joint project of the MIT Laboratory for Computer Science and the M. I. T. Library System, with the cooperation of the MIT Artificial Intelligence Laboratory. Most of the activity of the second level

An Informal Proposal For MIT Participation In

An Electronic Library Consortium Plan

6 June 1990

1. OVERVIEW

A small consortium of research universities, currently consisting of Carnegie Mellon, Stanford, the University of California at Berkeley, and MIT, is proposing to create a working, useful prototype of an online, electronic library consisting of the technical reports of their respective Computer Science laboratories. Each institution would create a network–attached server to hold scanned images of its own technical reports, and bibliographic records that describe the technical reports and that contain pointers to the images. The bibliographic records would be exchanged with each of the other institutions and placed in a catalog server at each site.

The purpose of the proposal is to obtain early experience with the distributed electronic library of the future, working with a database that is easily available, that has a minimum number of complications arising from such things as copyright ownership, that has a volume large enough to be useful and interesting, and that has an audience that is both motivated and likely to have access to appropriately powerful workstations to use the library from their offices. The complete consortium proposal can be found elsewhere. This document outlines briefly the plan for MIT participation.

II. MIT PARTICIPATION PLAN

MIT proposes two potential levels of participation, with corresponding budgets:

II.1. At the first level, MIT is involved at the minimum level of activity that is useful both to MIT and to the consortium, providing scanning of its computer science technical reports and making the consortium data base available to the MIT community. System development and installation would be handled primarily by imitation of the corresponding activities at CMU, with little or no innovation. The funds proposed consist on the hardware side of scanning equipment, a storage server, a catalog server, and a minimum number of workstations to be placed in library and reading room areas. Funds for personnel are also required, of two categories. The library and reading rooms currently are very low–technology activities, so technical support is needed to set up and run the electronic systems; in addition, staff time is required to do scanning, cataloguing, and feeding consortium catalog materials to the catalog server.

II.2. At the second level, MIT does all the things of the first level, and in addition undertakes to explore some of the organizational and technical opportunities that the consortium data base provides:

4. RESOURCES NEEDED BY OTHER UNIVERSITIES

To use this software, another university would need:

- An indexing server running Ultrix. The performance and storage requirements depend upon the volume of data to be stored, the number of users, and whether or not a separate image server is used.
- Unix workstations running the X/Motif window manager.
- Ethernet connections or equivalent connections on–campus and to the Internet.
- A license from OCLC to run the Newton database system.

The software has been developed on Vax computers. Conversion to DecStation computers is scheduled. The user interface will be supported on other Unix computers, beginning with the SparcStation, as resources become available. We recommend that each site have an experienced systems administrator available to support the software.

Architecture

- Database protection and authentication issues; statistical monitoring and reporting, special functions required by the user interface; desirable features for distributed software development (6 mm).

  Total architectural = 6 mm.

Distribution & Documentation

- Test portability of User Interface to non–Vax architecture — at least SparcStations and DecStations (4 mm).
- Packing and testing (8 mm).
- Document client/server programming interfaces to protocol (2 mm).
- Document interfaces to Hesiod and Kerberos beyond Athena documentation (2 mm).
- Documentation and toolkit for converting machine readable, marked up, technical papers to Newton format (4 mm).

  Total distribution and documentation = 20 mm.

3.3 Staff

These costs are what would be required to augment the Mercury team to carry out this work over 24 months, using some existing staff and hiring new skills.

Database & systems	$0	use existing staff
Architectural & Networking	34,000	add $17k/year to existing line
User Interface	76,000	new hire
Test & Packaging	56,000	new hire
Tech Writer	50,000	new hire
Subtotal salaries	216,000	
Fringes & indirect cost	199,000	
Total salaries	415,000	

Most of this software is already under development and many of the key components are in the advanced stages of testing. This section describes the extra work that is required to accelerate the development of software for bit–mapped and Postscript images, and to document and package software for use by other universities.

The project duration is 19 months after hiring has occurred. This consists of 13 months of development and 6 months of project field test. Software will be available for 3 workstations: VaxStations, DecStations and SparcStations. We will use optical servers from the Epoch Corporation. These NFS based optical servers will provide optical access to Newton retrieval servers via NFS.

3.2 Detailed Work List

All estimates are in man months (mm).

User Interface

- Add window to display bit–maps, bi–tonal and gray scale displays, including Fax G4 decompression (4 mm).
- Provide read–ahead for bit–mapped pages (1 mm).
- Add support to display Postscript documents (4 mm).
- New features for member universities (4 mm).

Total user interface = 13 mm.

Networking

- Confirm compliance of Z39.50 implementation (6 mm).
- Implement Database Information Server – Provides information for user interface to retrieve and interpret records at a remote database (3 mm).

Total networking = 6 mm.

Database & General Systems

- Modify Newton to handle potential long access time to optical media (2 mm).
- Tools for converting locally generated abstracts to Newton databases (2 mm).
- Tools for linking image data to abstracts (2 mm).
- Evaluate, acquire and install image server hardware (6 mm).

Total database and general systems work = 12 mm.

We are currently in the midst of a large workstation purchase which might result in a substantially lower price for workstations.

2.4 Backup Server

The backup server is used to back up the indexing server and to log changes to the image server.

	List Price	CMU Price
DecStation 3100	$16,000	
1.3 Gbyte disk store	10,000	
Dual 8mm tape unit	10,000	
Total backup server	36,000	26,000

2.5 Hardware Summary

	List Price	CMU Price
Total image server	$133,000	$133,000
Total index server	385,000	211,000
Total workstations	315,000	157,000
Total backup server	36,000	26,000
TOTAL	869,000	527,000

3. SOFTWARE DEVELOPMENT

3.1 Deliverables

This project provides network access to a technical report database on the Internet. The database consists of a single bibliographic database (which can be replicated), and, at each university, a database of scanned text images. The bibliographic database is collected from a number of sources but is provided in a common format to the development group. The image database is provided by an outside source and scanning is not included in these estimates. The software provided consists of:

- The Mercury User Interface using the X/Motif window manager
- The Mercury Z39.50, Kerberos, and Hesiod libraries
- The Mercury White pages and Yellow pages servers
- The Newton Text Retrieval System
- Utilities for maintaining the bibliographic and image databases

for experimentation in caching, pre–fetching, and other performance enhancing algorithms, with much larger databases than this project will generate initially.

The suggested configuration is an Epoch–1 optical server with 30 Gbytes of optical storage, with 650 Mbytes of magnetic cache, which can be extended to 150 Gbytes. The price quoted for this system is $133,000.

	List Price	CMU Price
Total image server	$133,000	$133,000

2.2 Index Server

The index server will be used to store the bibliographic records and the source text of documents. They will be indexed in the Newton database system. This server will also be used for database development and experimentation.

The configuration is as follows:

	List Price	CMU Price
Digital 5810 with 64 Mbyte	$132,000	
10 Gbyte storage	197,000	
Extra disk controller	26,000	
Extra Ethernet	5,000	
Extended warranty	25,000	
Total index server	385,000	211,000

2.3 Workstations

Workstations are required for use by software developers, by the library and publications staff, and in key locations within the libraries. These workstations will also be used for experimentation in user interfaces.

User workstations are DecStation 3100s with 19 inch displays, 16 Mbyte memory and 200 Mbyte disk. Developer workstations have 400 Mbyte disk.

	List Price	CMU Price
4 DecStation 3100 for developers	$79,000	
12 DecStation 3100 for library use	236,000	
Total workstations	315,000	157,000

Carnegie Mellon University

Mercury Electronic Library

Computer Science Reports Project

William Y. Arms

30 April 1990

1. OVERVIEW

This is to confirm that Carnegie Mellon wishes to be an active member of the Computer Science Technical Reports project. We plan to make the following contributions:

- To create bibliographic records of our reports, and to make them available to other members of the project.
- To build an online index of all the bibliographic records, and to make the index available to the Carnegie Mellon community (and others by agreement).
- To build an online collection of our reports in image format, which can be read over the Internet, and to make it available to all members of the project.
- To collect the source text of our reports for further experimentation.
- To complete all Mercury software which is needed for this project, document, package, and distribute it.

Our only restraint is resources. The requirements are hardware, and money for software development. If our full hardware requirements can be funded, it may be possible to transfer funds from other parts of the Mercury project to support software development, though not necessarily on the accelerated schedule that we hope for.

2. HARDWARE NEEDS

The following equipment list is ranked in order of priority, with the most urgent requirement first.

In the following tables, we show two prices, the list price of equipment and the approximate price that Carnegie Mellon would expect to pay if we make the purchases ourselves.

2.1 Image Server

The image server will be used to store images of documents is bit–mapped or Postscript format. We plan to use erasable optical disks for storage. This server will also be used

production. Meridian will continue to be available to serve in the role of coordinator. The Information Sciences Institute, another organization closely linked to DARPA, was invited to the workshop, provided valuable technical advice, and has been invited to participate as an initial user of the system and as a continuing technical advisor. DARPA itself remains convinced of the worth of the effort and will continue to assist the consortium in whatever ways it can.

Massachusetts Institute of Technology

Massachusetts Institute of Technology (MIT) proposes two potential levels of participation. The first level consists of full participation in the consortium, using the CMU software. At the second level MIT proposes to use the electronic library project to integrate the diverse report production and dissemination activities of its various laboratories and departments into a more unified and effective system. MIT also proposes to collect available source text and convert it into a common format for experimentation with full–text search. Proposed hardware, at the second level, consists of an image server, an index server, a backup server, 13 workstations, an experimental index server, a scanner, a fiche camera upgrade, and a high speed printer, for a total of $442K. Proposed staff, including overhead, and at the second level of effort, totals $540K for a 12 month period. The combined hardware and staff costs total $982K. Total computer science technical report production is estimated at approximately 12K pages per year.

University of California

University of California (UC) proposes participation in the project by the Division of Library Automation, a UC system–wide organization, and the computer science department of UC at Berkeley (UCB). Their proposal segments the costs by bibliographic and content (initially image) components. Within the bibliographic component, UC proposes two tasks in addition to providing records for UCB computer science technical reports. The first is to make bibliographic records from all consortium participants available to the overall Internet community through its MELVYL online catalog. The second is to create and provide to all interested parties programs to map between the text–based consortium records and the MARC format, which is universally used as a bibliographic data storage and transmission format among libraries. On the content side, UC does not propose any hardware, other than a scanner. UCB has access to a large archival storage system as part of the Mammoth project and this has already been interfaced to the Postgres database system. They propose to build on these existing components to investigate the use of Postgres for source document storage and access (both image and text), to investigate image server protocols and architectures, and to extend the project to complex electronic documents containing data sets and object code. Proposed costs, excluding overhead and covering a period of one year, total $32K for the bibliographic component and $132K for the content component, for a total of $167K, almost of which is for personnel.

Other Organizations

DARPA/ISTO has employed the Meridian Corporation to identify the potential members of the consortium, identify the issues, coordinate the preliminary discussion, organize the workshop, and facilitate post–meeting discussion and proposal

an enabling framework to explore different aspects of the electronic library. Hardware needs also vary, depending in part on what is currently available at each institution.

Carnegie Mellon University

Carnegie Mellon University (CMU) proposes to leverage their ongoing work in Project Mercury* to develop, document, and package a software suite for distribution to any present or future members of this consortium. This package, combined with appropriate personnel and hardware, would allow full participation in the consortium. The package would consist of:

- user interface using the X/Motif window manager
- Z39.50, Kerberos, and Hesiod libraries
- white pages and yellow pages servers
- Newton Text Retrieval System
- utilities for maintaining both bibliographic and image databases

Much of this software is currently under development. They propose a 19 month schedule to accelerate development, extend it to accommodate bitmapped and Encapsulated PostScript images, and to package and document the software for use by others.

Proposed hardware for CMU participation in the project consists of: an image server, an index server (for both the bibliographic data and for source text of full documents), a backup server, and 16 workstations. Total proposed hardware costs are $527K and total proposed personnel costs, including overhead, are $415K, for a combined total of $942K. Total computer science technical report production is estimated at approximately 10K pages per year.

Stanford University

Stanford University proposes to use the CMU software suite as it is developed. Their hardware requirements consist of an image server, four workstations, a high speed printer, a scanner, and some local network extension equipment. Proposed hardware totals $200K. Personnel costs for an 18 month period, including overhead, total $149K, for a combined total of $349K. Total computer science technical report production is estimated at approximately 17K pages per year.

---

* Project Mercury is a joint effort of CMU and the Online Computer Library Center in electronic library prototype building. It has been supported by DEC, the Pew Memorial Trust, and AAAI.

Use of standard protocols to link the various components of the system will allow each participant to choose its own database systems, workstations, user interfaces, etc. (Details of standard and protocol selection appear in the proposals.)

Expansion

The bibliographic data sharing will be opened up to all interested institutions as soon as possible. The bibliographic format must be thoroughly tested, however, before any attempt is made to use it as a national standard. It must prove effective both for data entry and for information retrieval. Expansion of image delivery to additional institutions, on the other hand, is not anticipated in the near term. This is a more sensitive area, for both authors and administrators, and deserves careful study.

Later Phases

Later phases of the project will focus on availability of full–text for searching, alternatives to bitmaps for full document display, and expanding both participation and the number and types of documents covered. There are a number of levels of full–text to examine, from completely unstructured ASCII to highly structured ASCII, e.g., SGML. The more semantically meaningful the text representation becomes, the more potentially useful it is for information retrieval and local navigation. Bitmaps of document pages, even with compression, are weighty items to move on demand across national networks, and more compact representation formats are highly desirable. These two aims are linked in the notion of compound documents. Scaling up the project, both in terms of material covered and participants, serves the dual purpose of providing improved access to computer literature and of pushing the development of local and national networks. These more ambitious goals require that the participating institutions collect document sources as well as bibliographic data and images. These sources, in whatever format they can be acquired, need to be banked for the future of the project. The project, especially with these sources in hand, should serve as an excellent testbed for all of the issues associated with electronic libraries, as well as immediately providing a badly needed resource.

THE PROPOSALS

The four proposals, initially submitted to DARPA/ISTO between April 30 and June 15 of 1990, all subscribe to the basic tenets given above. The institutions are agreed on generating and distributing the bibliographic records and on generating and making available for network access the corresponding image files. Within that framework, however, they propose varying amounts of activity for varying amounts of dollars. In some cases these varying strands intentionally mesh, e.g., sharing software development, while in other cases they intentionally separate, thus using the project as

Each institution will be responsible for the use and maintenance of its received copy of the data. There will be no restrictions placed on any institution's use or redistribution of the bibliographic data and abstracts.

The bibliographic format will be as agreed upon by the participating institutions. The goal is a simple record structure consisting of a small number of tagged fields. Certain fields will be deemed mandatory, i.e., the participating institutions will be expected to supply this data whenever it is available. A second set of fields will be deemed optional. A unique publication id scheme will be developed, most likely based on ids currently assigned by the producing institutions. The institutional source of the record must be included, as must a record type indicating whether the record is new or is a revision of an earlier record. Some institutions may wish to use this bibliographic data in the creation of catalog records, but this and related topics are outside the scope of this project and are left to each institution. Adherence to the specified format shall be the responsibility of each institution. Complaints or problems with non–adherence shall be directed to the coordinator at each institution.

There will be no attempt, at least initially, to mandate any sort of unified indexing or classification scheme. This does not preclude an optional field to hold this sort of data, nor does it preclude one or more organizations indexing or classifying any or all of the records and offering this work to others in the form of revised records.

Page Images

Each institution will provide page images, initially in the form of bitmaps, for those technical reports which it has announced in the bibliographic data discussed above. Unlike the bibliographic data, the images need not be replicated at each participating site, but will be available on local servers at each originating site. The bibliographic data will contain the pointers needed to locate the image data. The image servers at each location will operate under an agreed–upon protocol for responding to a request for one or more page images. Requests will initially be honored only for a tightly controlled group of potential requesters.

The bibliographic and image data will match, except for those cases in which the image of the full document has been withdrawn and the bibliographic record has been allowed to remain. An institution could decide, for example, to leave in place the bibliographic record for a technical report which has been substantially altered and now appears as a journal article, but to withdraw the image of the complete original report. In those cases, request for display of all or part of the complete document will be answered with an appropriate message referencing the current state of the document.

Original image capture will remain the responsibility of each institution. This does not preclude one or more institutions from subcontracting this activity to an outside service bureau.

goals, including expansion of the bibliographic component to the rest of the country and expansion of the full document component to direct input of and access to source material, in addition to images.

Computer science technical reports were selected as an ideal starting point because they will produce a database of literature that is now difficult to obtain or even be aware of in a timely fashion, would be useful to an audience that is both motivated and likely to have access to the needed hardware, and will raise the fewest complications from ownership issues.

Independence of external funding is one of the goals of the consortium, but external startup funding is required. In the long run, such a system would be less expensive to operate than the current approach of mailing bibliographies, ordering by mail or telephone, shipping paper copies, cataloging at each library, etc., but in the short run the two approaches will have to co–exist, an increased strain on already overburdened university library budgets. A relatively small amount of funding would be sufficient to energize this consortium. It is a significant opportunity.

THE AGREEMENT

A workshop on this topic was held March 29 and 30, 1990. It was sponsored by DARPA/ISTO and attended by representatives of the above universities plus ISTO, Information Sciences Institute, and Meridian Corporation The timeliness of the workshop was underscored by the relative ease and rapidity with which a consensus was reached. The meeting itself produced a basic agreement on a cooperative approach and that agreement was confirmed and expanded upon in an extensive email exchange following the meeting. This led to the four proposals, all of which are attached as appendices. The remainder of this section describes the basic agreement. It is followed by a summary of the four proposals.

Bibliographic Data and Abstracts

Each institution will be responsible for creating bibliographic records and abstracts according to a specified format. In practice, this is likely to be done at individual publication offices, i.e., the organizations responsible for publishing the reports. These records will then be sent, via the Internet, to each of the other participating institutions, thus allowing, although not requiring, the creation of a complete, replicated bibliographic database at each institution. The details for sending the records remain to be worked out, but it is agreed that submissions will be authenticated, there will be a single data feed to each institution, and a single office or person will be responsible for the accuracy of the mailing list. Electronic mail (probably using a mail reflector) has been proposed as the initial distribution mechanism. As the project scales up, however, it is possible that this method will require re–examination.

National Technical Report Repository

A Proposal for Electronic Library Development

The vision of the electronic library has been with us for over forty years. Increasing amounts of detail have been filled in over that period, but the basic vision has remained the same — provide researchers, practitioners, and students with needed documents directly at their desks, via an electronic network. Recent technical advances, e.g., high–speed networks, large disk storage, powerful yet relatively inexpensive workstations, and efficient image compression algorithms, have brought us to the point where it is now possible to proceed with functioning prototypes of this vision. Such prototypes could provide immediate benefits to researchers, et al., while also serving to illuminate areas requiring further work. Convenient and timely access to technical literature could energize the research community and maximize return on research dollars by minimizing redundant research and encouraging rapid exchange of ideas between research centers. While difficult to quantify, it is clear that even a small improvement in researcher efficiency would return the startup costs of such a program many times over.

The purpose of this document is to introduce and provide a common backdrop for four top–level proposals on the topic of electronic library development. Carnegie Mellon University, Massachusetts Institute of Technology, Stanford University, and University of California propose to form a consortium focusing, initially, on providing distributed electronic access to computer science technical reports.

The initial plan revolves around a merged bibliographic database, replicated at each site, with links to distributed page image files containing the complete documents. This would allow users, at their workplace, to search through an absolutely current database of technical report citations and abstracts and to receive images of the complete documents, on request, in real or near real–time. The consortium has numerous other

National Technical Report Repository

A Proposal for Electronic Library Development

July 1990

Containing Proposals From

Carnegie Mellon University

Massachusetts Institute of Technology

Stanford University

University of California

Prepared by

The Meridian Corporation

For further information contact
Laurence Lannom or Richard Lessar
(703) 998–3600