The Computer Science Technical Report (CS-TR) Project: Considerations from the Library Perspective

Greg Anderson (MIT)
Rebecca Lasher (Stanford)
Vicky Reich (Stanford)

June 26, 1996

This technical report is also available from Stanford University as Computer Science Technical Report CS-TR-95-1554

A revised version of this report was published in The Public-Access Computer Systems Review 7, 2 (1996), an electronic journal of the University of Houston Libraries. That version is available on the World-Wide Web at <http://info.lib.uh.edu/pr/v7/n2/ande7n2.html>.

The Computer Science Technical Report (CS-TR) Project: Considerations from the Library Perspective

Greg Anderson (MIT)
Rebecca Lasher (Stanford)
Vicky Reich (Stanford)

Abstract

In 1992 the Advanced Research Projects Agency (ARPA) funded a three year grant to investigate the questions related to large-scale, distributed, digital libraries. The award focused research on Computer Science Technical Reports (CS-TR) and was granted to the Corporation for National Research Initiatives (CNRI) and five research Universities. The ensuing collaborative research has focused on a broad spectrum of technical, social, and legal issues, and has encompassed all aspects of a very large, heterogeneous distributed digital library environment: acquisition, storage, organization, search, retrieval, display, use and intellectual property. The initial corpus of this digital library is a coherent digital collection of CS-TRs created at the five participating universities: Carnegie Mellon University, Cornell University, Massachusetts Institute of Technology, Stanford University, and the University of California at Berkeley. The Corporation for National Research Initiatives serves as a collaborator and agent for the project.

As the project comes to a close, accomplishments include: a large digital collection; an exchange format for bibliographic data (RFC1357 superseded by RFC1807); a distributed, web-based delivery protocol (Dienst); an information awareness service (Sift); an approach to interoperability (Kahn/Wilensky paper); and a web catalog tool

---

Anderson was MIT Libraries Associate Director for Systems and Planning at the time this technical report was written. He is now MIT Information Technology Discovery Process Leader (ganderso@mit.edu)

Lasher is the Head Librarian, Mathematical and Computer Sciences Library, Stanford University (rlasher@forsythe.stanford.edu).

Reich is Assistant Director Highwire Press and Information Access Analyst, Stanford University Libraries, Stanford University (vicky.reich@forsythe.stanford.edu).

(Lycos). Perhaps the most enduring accomplishment of the project however, is the mutual respect that has grown between the computer scientists and the librarians who are working together to investigate the challenges of electronic library information.

This technical report summarizes the accomplishments and collaborative efforts of the CS-TR project from a librarian's perspective; to do this we address the following questions:

1. Why do librarians and computer scientists make good research partners?
2. What has been learned?
3. What new questions have been articulated?
4. How can the accomplishments be moved into a service environment?
5. What actions and activities might follow from this effort?

Contents

I. Introduction:

Be favorable to bold beginnings

--- Virgil

The ARPA sponsored CS-TR project is one of the earliest sustained investigations into the system engineering of digital libraries. The notion of a digital library based on Computer Science Technical reports began as a somewhat pragmatic enterprise, but as more fundamental questions and opportunities arose, the project grew into a large-scale effort that has pioneered collaborative research. This prototype library is being used to investigate basic questions around building, managing and accessing networked, interoperable collections of valuable intellectual property. The project's main accomplishments can be summarized:

1. Librarians and computer scientists are good research partners.
2. The project has created a prototype service.
3. The critical issues associated with the evolving concept of digital libraries have been better articulated through practice and deeper research.

Research Partners - If we accept that we are living in the Information Age and that a challenge for this age is to give people tools with which they can successfully use networked information, then librarians and computer scientists are natural collaborators to address this challenge. Computer scientists and librarians each bring to the discussion complementary technical skills and perspectives. Computer scientists have a large view of the network, new approaches to information retrieval, and an openness to change. Librarians have content, and a historical, enduring view regarding service and responsibilities for our intellectual heritage. Both communities share the academic values of sharing openly and the desire to foster the creation of new more powerful knowledge. In this project, the librarians have benefited from the computer scientist's cultural value of exploration and learning by doing. The computer scientists have benefited from the librarian's broad perspective and integrative skills. The coupling of content and carrier, scale, inter-operability, and mutual respect for professional knowledge and abilities has served to create a productive, dynamic atmosphere.

Prototype Service - The project testbed supports both service and ongoing experimentation. While the prototype service is available now for public use, the testbed and its services are also continuously changing. This CS-TR project highlights the tension between providing reliable services while experimenting with new capabilities. Moving into the future and contributing to new arenas of digital information while maintaining perspective and providing daily services are challenges for individual librarians and innovative library organizations. In the CS-TR project, librarians have continuously examined the long term viability of the effort. At each stage of the effort, it has been important to remember the research nature of the project and that digital libraries are in their nascent state. Whatever we build today will be superseded by more powerful knowledge and services in the future.

Issue Articulation - The investigation and better articulation of our early research questions provide the forum and starting point for solid achievement and greater progress in the future:

- How do we build technologies that make scholarship more effective?
- What do we really mean by a digital, virtual library? The technological, educational, social, economic, and legal questions that we have articulated in this project are fundamental to the networked environment. As the project comes to close in 1995, it is important to link and transmit our learning to other digital library pioneers. An initial contribution we can make is to impart a sense of humility given the scale of the issues.

I.1 History

Discussions for the CS-TR project began in 1990 and evolved finally into the structure in place today. The original question posed for the project was straight forward: how can we make computer science technical reports more accessible to researchers? Computer Science Technical Reports are an important body of knowledge, they are often difficult to locate because they are normally published by the academic/research departments, and we believed that the intellectual property issues were not terribly complex. Through the early discussions among the participating institutions the horizon of the issues expanded and this broadened view was presented to potential funding agencies. With ARPA funding in 1992 and CNRI's role as contract administrator, it became apparent that we had the potential to set the pace for several important pieces of the digital library: distributed, virtual collections spread across the network, development of sophisticated linking mechanisms that would enable the location and retrieval of information no matter where located, incorporation of mechanisms to handle intellectual property issues in a digital environment, and finally, better understanding of the service and scholarly productivity issues for electronic library services. The consortial arrangement of the project has enabled each institution to pursue separate but linked approaches to these issues. Each of the five participants has placed its own TR's online at its home location. Through network based searching and retrieval mechanisms, we have explored the issues involved in sharing, rather than duplicating, on-line information. This sharing has created an early prototype of a virtual collection of Computer Science Technical Reports and serve as a model for building similar virtual collections in other areas.

The research goals of the project varied with each participant. In A Proposal for M. I. T. Participation in an Electronic Library Plan (10 November 1992), however, most of the key points involving technical, organizational, service, and data questions are enumerated:

1. to obtain early experience with a core function of the distributed electronic library of the future,

2. to work with a database that is readily available, that has a critical time-sensitive value, and that is already well-known and valued by its target audience,

3. to explore the architecture, design, and work-flow issues associated with making information available in digital form,

4. to work within the research/prototype domain with a volume of information large enough to be useful and interesting and that can scale to an operational system,

5. to provide an important service to an audience of researchers, faculty, and students who are motivated and likely to have access to appropriately powerful workstations to use the library from their offices.

Each campus has pursued research questions within the framework of these goals. CNRI has led the coordination, discussion, and facilitation of the individual efforts and has contributed its own research on linking mechanisms and electronic copyright management. In sum, the project has enabled investigations into digital libraries on a number of facets that have yielded substantive results.

I.2 Basic Design

The project's core design is based upon the construction of a bibliographic records database that describe the TR's and enable linkage to the page images of those TR's. The concept of the database has been debated over the course of the project, should it be centralized and replicated at each site, or should it be distributed where each site maintains the index record only for its own collection? The nature of the linking mechanism between the record and the images has been a topic of lively discussion and development. We must assume that the TR bibliographic record will be stored in a different location from the page images and that both the records and the images may move to other machines during their lifetimes. What linking mechanism will support this location flexibility and maintain high, efficient, performance?

In addition to images, project staff also experimented with the full text of the TR's, obtained from the source files of the TR or through OCR techniques on the images. Together, these files will enable exploration and evaluation of: full text retrieval mechanisms; data integrity for huge stores of data; and citation linking of references across documents (for example a link from a footnote or citation in one document to the cited document itself).

II. Investigations:

This section focuses on the collaborations among the CS-TR participants. A great deal of research was done by the individual institutions that is not mentioned in the body of this report. Detailed descriptions of these activities can be found on each University's web page; these are all linked from the URL: http://www.cnri.reston.va.us. A list of the products can be found at the end of the report.

II.1 Bibliographic record format

Many computer science R&D organizations routinely announce new technical reports by mailing (via the postal services) the bibliographic records of these reports. This paper alert service has some obvious drawbacks: mailing costs; postal delays; the format is not amenable to convenient filing for later retrieval and searches. The CS-TR participants wanted to move from paper to electronic sharing of bibliographic records. To accomplish this task however, we needed to all use one bibliographic record exchange format.

The group discussed alternatives. We wanted a simple format, for people and for machines; one that was easy to read ("human readable") and easy to create. (These bibliographic records are usually produced by secretaries or publications coordinators). We knew we were possibly choosing an interim format as automatic and full-text indexing methods may supersede bibliographic records.

Using USMARC (US Machine Readable Cataloging), prevalent in library cataloging process, was considered early in the project and discarded. USMARC is very complex, is not easily taught, nor is it accepted by non-catalogers. Project staff were concerned that the complexity and the high level of training necessary to catalog in USMARC may cause significant time delays between TR publication and bibliographic record. For this CS-TR project, the possibility of a delay was unacceptable.

The Consortium came to agreement on naming authorities for institutions but beyond that no standardization rules like AACR2 were discussed.

BibTeX and Refer were also considered and rejected. Neither had the required CS technical report fields (like Computing Reviews Category, monitoring, funding, contract organizations, or grant number).

CS-TR participants created their own bibliographic format: 'RFC1357, A format for mailing bibliographic records' later superseded by RFC1807, 'A format for bibliographic records'. Basic design principles of the RFC 1357/RFC1807 were:

1. identification and creation of data elements basic to citation creation, management, and retrieval.

2. creation of the bibliographic record must coincide as closely as possible with the publication of the TR.

3. creation of the RFC1357/RFC1807 record should be possible via machine parsing of CS-TR title page data and/or by staff in the CS-TR publications operation, not by library catalogers.

4. to provide core information for the more formal library catalog record, RFC1357/RFC1807 UCB, Stanford, and Cornell built translators that map into other formats, including USMARC.

Once the group decided to create a new format, the development and implementation of the format proceeded quickly. We were not constrained by the older formats and could add fields as desired.

II.2 Centralized vs. Distributed Indexes

Once the bibliographic record format was created the discussion turned to centralized vs. distributed indexes. Long conversations ensued where the participants argued the virtues, value, and scalability of centralized and/or decentralized indexes for very large distributed collections.

One of the early goals of the project was to develop an inter-operable, distributed collection whereby each site would develop its own testbed architecture, create consistent content based on the G4-tiffb standard, and then experiment with interoperating and sharing those collections across different systems. In the end no conclusions were reached and the above goal was not met. We know that neither centralized nor decentralized servers will scale. Eventually a more complicated, yet to be determined, architecture may emerge which will involve replication of an institution's indexes on several servers around the country. This effort will require more research and a lot of cooperation between insitutions.

In order to get started, Cornell developed Dienst, which is a protocol and implementation that provides Internet access to our distributed collections. The indexes are produced and kept at each institution. Each institution is required to run the Dienst server protocol. Dienst does permit the "single distributed collection model" but it is not an inter-operable model running on different software and server platforms; For a description of the newest version of Dienst see http://www.ncstrl.org/Dienst/htdocs/Info/protocol4.html.

In the Dienst architecture there are four classes of services. A Repository Service stores digital documents, each of which has a unique name and may exist in several different formats. An Index Service server searches a collection and returns a list of documents that match the search. A single, centralized Meta Service (also called a Contact Service) provides a directory of locations of all other services. Finally, a User Interface service mediates human access to this library. All these services communicate via the Dienst protocol. [Dienst Web page at http://www.ncstrl.org]

A group of sites sharing the Dienst protocol form a single distributed collection. Each site will typically run repository, index, and UI services for documents issued by that site. One of the sites will run a Meta service, thus defining the set of sites that make up the collection. [Dienst Web page at http://www.ncstrl.org]

> From the standpoint of a Dienst user, a document collection consists of a unified space of uniquely identified documents, each of which may be available in a variety of formats. Using publicly available World Wide Web clients, users may search the collection, browse and read individual documents in any of their available formats, and download or print a document (Davis, 1995).

In the current implementation of Dienst, CS-TR users can query all or selected institutions using combinations of keywords in fields (author, title, etc.). The search is performed in parallel at user selected sites. If a server is unavailable the search will time out and display a message to the user that a particular server is down. Some institutions have implemented full-text searches but those searches are limited to single institutions.

> There are currently approximately 14 CS-TR Dienst servers. At any given time it is likely that at least one of the servers will be down. Further work needs to be done in

two areas: begin replicating index servers to increase availability and response time; add persistent search which continues to attempt to contact non-responsive sites (Davis, 1995).

II.3 Images vs. SGML vs. Postscript vs. ASCII

The pros and cons of a standardized format (images, SGML, Postscript, ASCII) for the technical report documents was vigorously debated; the outcome? Tiff-b image format (also called Group IV FAX compression in Tiff format) was selected as the project standard. This decision was supported by at least the following factors: in 1992, image formats were standard and many commercial software packages were available on multiple platforms; retrospective paper reports could be converted to image format; project participants were eager to populate servers with both retrospective and prospective reports; and researchers did NOT want to spend resources on document mark up, document conversion or on developing new standards. Two faculty members of the consortium believe that images should remain the ultimate version of record because they provide the simpliest exact representation of the document and can be exported to new software and platforms over time. In brief, the consortium chose to try and populate the architectures with content rather than trying to solve the issue of how to format the content.

Many of the CS-TR institutions have made multiple formats available on their servers. All formats are available through the Dienst protocol. Image is a requirement for the project but most institutions also offer Postscript and ASCII, particularly for the newer reports.

II.4 Scanning and OCR

A lot of investigation and sharing amongst the collaborators was done regarding scanning and OCR in order to purchase equipment and software. Although there was no dpi requirement, the group agreed in general to scan at 300 dpi or greater because less might require rescanning as more sophisticated systems were developed. Each institution purchased different equipment and software. As long as tiff-b image was the end product we did not need the same equipment. In fact, the research encouraged different implementations.

MIT did the most research on production scanning, archiving, and recordation of the scanning process.

The MIT Library 2000 testbed effort focused significant attention on production scanning. This emphasis is based upon the hypotheses that scanned images of documents will be an important component of any future electronic environment. At its core, the digital library must contain high quality content, and, for the foreseeable future, much of that content will come from the conversion of paper format information to scanned images. Further, the creation of a large corpus of quality information provides the testbed content for investigations into system architectures, electronic information management, retrieval, and long-term storage. Basic principles of the MIT scanning effort include:

1. Materials should only be handled once. The design of the scanning environment should strive to achieve the greatest advantage in terms of price, performance, and quality. Libraries and publishers cannot afford to re-scan materials as technological capability increases. For the original paper artifact, scanning once is also preferable. To adhere to this principle, good paper workflow, management, and content selection is important.

2. Scanning should capture as much information as possible in the single scan principle. Current technology cannot exploit all of the bits captured, future technologies, however, will be able to exploit all nuances of the captured information. The MIT scanners are capable of a resolution of 400 pixels per inch, with eight bits of gray-scale per pixel. These create very large files (about 16 Mbytes per scanned page), which are rendered down to the agreed-upon interchange format for the project: 300 dpi, one bit per pixel, in CCITT Group IV FAX compression in TIFF format.

3. Quality control is critical. In order to achieve the first two principles, quality control methods must assure a high degree of integrity and confidence in the production environment. The MIT Libraries' Document Services has adapted procedures from its micro-reproduction heritage for this new production scanning effort. Document Services is using test targets from the Association for Information and Image

Management (AIIM) and the Institute of Electrical and Electronic Engineers (IEEE) to test calibrations on the scanner. Quality control is checked via file checksums and visual review of selected images.

4. Context of the images is important now and in the future. Because the underlying technologies will change and improve in the future, the CS-TR scanned images must provide enough context for humans and machines to understand both their content and structure in order to use them effectively. The MIT scanning effort has created a metadata record to provide information about the scanned document and the environment in which it was created. This record specifies both the form and content of the information that must be captured when a document is scanned, and becomes a component of the scanned form of the document. The record assists in viewing, displaying, or printing the image correctly; in understanding how to interpret the image, and in meeting contractual or legal requirements.

As a final note, this scanning effort required integral coordination and collaboration between an operational unit of the library and the computer science research group. The array of investigations, findings, and new questions have opened new paths for ongoing work.

II.5 Distributed Digital Object Services

Perhaps the most important intellectual discussion of the CS-TR effort has been the discussions on infrastructure and architecture for a large distributed digital library. The outcome of these discussions is captured in the Kahn/Wilensy paper titled "A Framework for Distributed "Digital Object" Services (Kahn 1995, http://www.cnri.reston.va.us/home/cstr/arch/k-w.html).

> This document describes fundamental aspects of an infrastructure that is open in its architecture and which supports a large and extensible class of distributed digital information services. Digital libraries are one example of such services; numerous other examples of such services may be found in emerging electronic commerce applications. Here we define basic entities to be found in

> such a system, in which information in the form of 'digital objects' is stored, accessed, disseminated and managed. We provide naming conventions for identifying and locating digital objects, describe a service for using object names to locate and disseminate objects, and provide elements of an access protocol (Kahn 1995).

The most important concept in the Kahn/Wilensky paper is the creation of a "handle" or a permanent unique identifier for every document. The "handle" is used to name the document on a server. A mechanism called a "handle server" maps the permanent unique identifier to the machine address. A working prototype of the "handle server" is available at CNRI and the "handle" functionality is being integrated into WWW browsers. Once browsers, like Netscape and Mosaic, know how to deal with "handles" a user/client with a unique identifier will be able to send a message to the "handle server" that will know on which document server the document resides. The client will then go to the document server using the URL or machine address. Unique permanent identifiers that are known worldwide and automatically map to the machine address using "handle servers" will be a very powerful tool for digital libraries. No longer will Web servers contain false links because the "handle servers" can update on a nightly basis.

The "handle" concept seeks to separate naming issues from location/address issues. Handles are not URL's; handles are an approach to a large-scale problem of naming objects that may change location over time. For libraries it is an important strategic and intellectual advance to be able to distinguish the name of an object from its address. For more information on handles and the handle system see http://WWW.CNRI.Reston.VA.US/home/cstr/handle-intro.html.

This design, and implementation of a network-wide enterprise to name and locate digital objects in the CS-TR project will have far-reaching ramifications for all digital library services. If libraries are to move beyond their physical walls and campus boundaries, and if libraries are to leverage the power of the distributed information base of the network to enrich services for their local community of users, a basic architecture for naming, locating, and accessing information must be well-understood and adopted.

II.6 Copyright

Intellectual Property is a fundamental issue in building digital libraries. At the beginning of this project, participants assumed that there would be few or no copyright issues with the Technical Reports. They assumed that the reports published at their schools were either in the public domain or that the rights were held by the University. Later, as questions arose, the group assumed that a single strategy would work for each institution. This proved to be naive. Upon investigation with legal counsel, researchers discovered that each school treated it's intellectual property differently and so five different approaches evolved.

At Stanford, the librarians took on the role of insuring that these IP challenges did not pose a risk to the University or to the Faculty. We identified scenarios that needed attention and began to meet with legal counsel to determine appropriate responses. These efforts helped us to articulate a set of intellectual property management models now used by the CS-TR projects at Stanford and Cornell. We encourage other schools to use these points to form guidelines for themselves. This is NOT legal advice. Every institution must rely on the advice of its own counsel. The worldwide legal environment is undergoing change and our current approach may become obsolete in the face of new laws and treaties.

• At most U.S. academic institutions, the author owns the copyright to any books, articles, or technical reports. Works published prior to March 1989 without a copyright notice are in the public domain (unless steps were taken within 5 years to establish copyright ownership). Works, with or without a copyright notice, published after March, 1989, are copyrighted.

• In most cases, reports that are produced, or that report on work sponsored by the government are not in the public domain. The government can make copies but at most institutions, the author owns the rights.

• Most CS-TR institutions ask authors to sign a form granting the institution non-exclusive, revocable, royalty-free license to publish, perform, display, and distribute the works. One author's signature is binding on multiple authored works.

• If an author has signed or plans to sign an exclusive agreement with a publisher for a particular work (or for substantially the same work) in a particular format, that author cannot then sign a non-exclusive agreement with the institution for the same work in the same format.

• If an author signs a non-exclusive agreement with an institution for a technical report and then decides to publish the same work elsewhere, the author should inform the publisher of this previous agreement. The author should then grant the Server Management written permission for the non-exclusive rights to publish, perform, and display the works before any works are loaded. If the author indicates s/he has already signed an exclusive agreement with a publisher, the technical report should probably not be mounted on the server without permission of that publisher.

• At some institutions the authors do not own the rights to their works. Each institution should be clear about copyright ownership before mounting technical reports on servers. The CS-TR group did not address the issue of third party rights in technical reports. When authors sign agreements it is assumed that the entire work is original or that the author has the rights to include non-original tables, charts, figures, etc.. This is one area that could be pursued by asking authors specifically about the originality of their works.

• There are several ways to manage technical reports that are submitted to publishers as articles.

Ask the authors not to sign the exclusive agreements with publishers. Ask them to modify the publisher's standard agreement to allow the institution to keep the work up on a server.

Make special arrangements with the publishers so the technical reports can stay on the servers even if an article is published.

If the author requests, remove the technical report from the server and point to the printed article.

• Include a notice with the technical reports to inform those viewers of their rights: for example, viewing on-line and transmitting over the network (this may well be legally considered a "performance" of the work); making printed copies; distributing copies to others; and selling copies to others. This relieves the users of guessing what restrictions might apply. Most likely the user will properly assume fair use restrictions apply, view the work and perhaps make a personal copy. But can they legally send a copy to a colleague? Cornell has chosen to clarify these issues by explicitly sub-licensing rights to the user (see http://cs-tr.cs.cornell.edu/). This makes clear the user's rights to make copies, quote, or redistribute the technical report. However, the sub-license must preserve the author's right to withdraw the technical report from further distribution.

III. Tension Prototype vs. Production

The CS-TR collaboration consisted of long discussions and compromise which created a system that is more logical than it would be without the collaboration. However, collaborations of this kind create tensions and what Leigh Star refers to as double bind situations (Star, 1995). Each institution was funded mainly to do research in specific areas of digital libraries. Each institution wanted to populate their servers quickly in order to get on with the research. In addition, all researchers wanted to get their systems used by as many people as possible. The products of the individual institutions and the collaborations have been quite successful. Lycos has thousands of accesses every day. Sift also has over 10,000 subscribers. The Dienst CS library now has 14 institutions using it as the production system to disseminate their technical reports. The prototype system is now being used in production. But now, enhancements and changes to the system are problematic; each time the Dienst code is upgraded, individual institutions have to work to implement the new system.

A dynamic outcome of this tension between research/prototyping and operations is the momentum to address the research questions embedded in this topic. For example, there are key research questions regarding distributed scale and linking of digital objects, the cognitive efforts to identify and present coherent collections to users, and the integration and evaluation of services that should be effective for the content and user regardless of media.

Even as we swing the pendulum from research to operations, the CS-TR Library will be used as a testbed for experimentation and continued research in Digital Libraries. The conclusion of the CS-TR project will offer several models for the migration of results from the research domain into library operations.

IV. Collaboration

Since the inception of the CS-TR project, librarians have worked in a collaborative atmosphere with computer scientists. Both groups of participants brought strengths to the project, and the cooperative results are superior to those if either group had conducted the project alone. Through ongoing discussions and consideration of common problems, such as the proposed handle mechanism, an atmosphere of trust and respect was created. The computer scientists were respectful of the librarian's concerns regarding the ongoing sustainability and operation of the service when the project funding ended, and the librarians gained greater insight and admiration for the innovation and "can do" spirit of the scientists.

For example, early in the design stage of the project, the appropriate structure and creation of the index record for the TR's was a key discussion topic. The computer scientists had expectations for fast and easy creation by a variety of staff or by technology, and librarians made the point of consistent record content and the flexibility of multiple uses of the index record. The result, RFC1807, accommodated both requirements in a sustainable, scalable manner. The records can be created by publishing assistants and can be created immediately upon acceptance of the TR. The records are also distinguished by consistent definition and use of the record fields, and conversion routines are in place to facilitate MARC record creation or use of the record in other formats. Another example is the collaboration of document service staff in the MIT Libraries with researchers to create an operational scanning service to convert TR's to page image form and to create the process and mechanism to accommodate massive amounts of information.

As the project comes to an end, each participating site is working through the issues of moving the research project into an operational service. This topic was first presented by librarians at the CS-TR meeting at Stanford in February 1993. Since that time, librarians have begun to share information on the process and issues of making this a sustainable, expanding project. Through efforts such as NCSTRL, librarians have served an educational and leadership role in ensuring that these materials will continue to be accessible and that the service is scalable and renewable. This ongoing respect and effort shared between computer scientists and librarians will continue to bear fruit long after the project has officially ended.

V. Expanding the CS-TR Project

The CS-TR project has generated substantial interest from other digital library efforts. We are now ready to encourage other institutions to use some of technology we have developed; to extend the current CS-TR library into a self-sustaining consortia effort that includes most major CS departments. For institutions that might want to participate, see, "How to participate in this distributed library" at http://cs-tr.cs.cornell.edu/Info/startupkit.html. This kit includes information on Dienst, the RFC 1807 and copyright considerations.

In particular, we are releasing a common bibliographic format and a document server, accessible through the World Wide Web, that allow searching, browsing, and displaying of technical reports from any WWW browser. If you are interested in participating, you should consider the following qualifying criteria:

1. Participating sites are required to adopt, implement, and use:
    - RFC 1807, the project's standard bibliographic record format
    - Dienst, a protocol and server for a distributed digital document library.

Adoption of these tools allow the site to automate the collection, management, and network availability of its own repository of computer science technical reports. And, the institution's report collection will become part of an expanding distributed library of technical reports through interoperation with other cooperating sites.

2. Doctoral granting U.S. institutions in computer science are invited to participate. Other institutions of higher education or commercial or government research laboratories who wish to participate should contact Rebecca Lasher, Computer Science Librarian (rlasher@forsythe.stanford.edu) to inquire about their possible involvement.

3. Before beginning to participate, institutions should evaluate their resources and commitment to this project. It is anticipated that this project will continue as an ongoing, operational service which will expand in content and participants even after the CS-TR project concludes. Therefore, institutions should only join if they feel they will be able to maintain their commitment over the long term.

At the June 1995 CS-TR meeting, the group agreed to ask the Computing Research Association (CRA) to endorse and to encourage the proliferation of this technology. A new consortium effort called NCSTRL or Networked Computer Science Technical Report Library which is a merging of two earlier systems, the (ARPA)-sponsored CSTR project and WATERS (Wide Area TEchnical Report Service) which was sponsored by the National Science Foundation (NSF). This new effort will continue to contribute to the broader Digital Library community. For more information see http://www.ncstrl.org.

VI. Observations

Over the three years of the project, every participant has gained a better understanding of the intellectual, organizational, social, and legal complexities embodied in a library. Building new, better digital services while preserving the endurable values of a library is hard.

Among the lessons learned are:

You cannot proceed far down the path of the digital library without facing very difficult issues: scale, content, use, intellectual property.

The underlying foundation of the digital library is content, structure, and organization. This foundation must be durable but flexible enough to be accommodated in future environments.

We should be creating good content the first time we try; we cannot afford to re-do the digital library with each new iteration of system design or access method.

Interfaces/interactions with the content will vary over time and may not be familiar to the digital library. Therefore a focus on openness and interoperability is critical.

The digital library is integrally involved with the nature of public and scholarly communication and information formats.*

The digital library is integrally involved with the economic and political environment within which information is created and sought.*

*(Adapted from: Sarah M. Pritchard, "Librarians: Real Expertise for a Virtual World," _Library Issues: Briefings for Faculty and Administrators_, vol. 15, no. 5, 1995).

VII. Conclusion

Libraries are operational, production oriented service organizations. The librarian's evaluation of research tends to focus on how successfully the products of research are integrated with or replace existing services; and how well they can be supported and renewed in a production environment. The CS-TR project has built several new prototypes, and they must now be extended into a production environment. It may be useful to think of the CS-TR project as beginning to address some of the key investigations for system design processes:

1. Discovery: matching the technology with the service vision. 2. Delivery: nurturing and developing this match in a prototype atmosphere to examine its feasibility and readiness for implementation. 3. Service: the ongoing operations of the service; continuous improvement of the service. 4. Support: provision of assistance, documentation, training, etc. 5. Integration: fit of the new service with the organization's overall architecture and services.

There are components of each process embedded in the CS-TR project. The project has made the most progress in the areas of Discovery and Delivery. The architecture and system engineering discussions in the area of integration have long-term outcomes for digital libraries. More precise questions for each process have been articulated. New efforts in the digital library arena will benefit from the accomplished work of the CS-TR project.

The results obtained by the CS-TR consortium provide a model of a working distributed digital library. These results will be useful for launching the new Joint Initiative DL Projects and as the conceptual frame work for further research. Beyond the current CS-TR effort, we believe that the CS-TR Consortium could also continue to contribute to the broader Digital Library community (Lynch, Garcia-Molina, 1995).

From the librarian's perspective, the CS-TR project offered the opportunity to work with and contribute to a world-class effort to transform scholarly communication. The learning experience was intense and gratifying. More questions have been formulated than were answered, but the new questions are better articulated and understood. The foundation laid by the CS-TR has immediate benefits and long term viability. We should note, we continue to evolve a definition of digital library. One of the questions is whether "digital library" is a real library - as we might define a library today - or whether the phrase is a metaphor for something entirely different. This report is a small step towards publicizing and presenting these CS-TR findings for broader dissemination and discussion in the Library community.

VIII. Products of the CS-TR project

Several public systems have been implemented with support from CS-TR and are available for public use. (Some of these services are under development and subject to change at short notice.) More information about these products are available on the respective institution's server. All CS-TR home pages point to each other. A good place to start is the CNRI server at http://www.cnri.reston.va.us.

• Dienst, a distributed search system for technical reports (Cornell) see http://www.ncstrl.edu for the information on the latest version of Dienst.

• Mercury, a centralized search system for technical reports (Carnegie Mellon) see http://rose.mercury.acs.cmu.edu/cstr_db.html.

• GLOSS, a system to help find relevant data sources (Stanford) see http://gloss.stanford.edu.

• SIFT, a system for performing wide-area information dissemination on USENET News groups and computer science technical reports (Stanford) see http://sift.stanford.edu.

• Lycos, a catalog of the Internet (Carnegie Mellon) see http://lycos.cs.cmu.edu.

• A handle server to maintain unique identifiers to objects in the Digital Library (CNRI) see http://WWW.CNRI.Reston.VA.US/home/cstr/handle-intro.html.

IX. References

Carnegie Mellon University = http://rose.mercury.acs.cmu.edu/arms/cnri/cmu-cstr.html

Cornell University = http://cs-tr.cs.cornell.edu/

Massachusetts Institute of Technology = http://cstr-www.lcs.mit.edu/cstr-www/

Stanford University = http://elib.stanford.edu/

University of California at Berkeley = http://cs-tr.cs.berkeley.edu/

We acknowledge the Annual and Quarterly reports from each institution.

'Start-up KIT' for CS-TR = http://cs-tr.cs.cornell.edu/

Davis, James R., Carl Lagoze, and Dean B. Kraft 1995. "Dienst: Building a Production Technical Report Server." Advances in Digital Libraries. ADL95 conference held in McClean, VA., May 15-17, 1995. Proceedings to be published by Springer-Verlag. Also see http://www.ncstrl.org for a description of Dienst 4.0.

Kahn, Robert and Robert Wilensky 1995. "A Framework for Distributed 'Digital Object' Services." version 5.3, dated 5/13/95. Available on the CNRI server at http://www.cnri.reston.va.us.

Lynch, Clifford and Hector Garcia-Molina. "Interoperability, Scaling, and the Digital Libraries Research Agenda: A Report on the May 18-19, 1995 IITA Digital Libraries Workshop, August 22, 1995."
http://www-diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html

Star, Leigh 1995. "Steps toward an Ecology of Infrastructure: Borderlands of Design and Access for Large Information Spaces", Susan Leigh Star and Karen Ruhleder. Submitted to Information Systems Research, Special issue on Organizational Transformations, edited by JoAnne Yates and John VanMaanen. Draft of March 4, 1995.

Acknowledgments

This work was sponsored in part by the Corporation for National Research Initiatives, using funds from the Advanced Research Projects Agency of the United States Department of Defense under CNRI's grant No. MDA-972-92-J-1029. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsement, wither expressed or implied, of ARPA, the U.S. Government or CNRI.

APPENDIX

RFC 1807 fields

For the entire RFC1807 see http://ds.internic.net/rfc/rfc1807.txt.

```
Request For Comments: 1807                          R.Lasher
Obsoletes: 1357                                     Stanford
Category: Information                               D. Cohen
                                                    Myricom


RFC 1807   A Format for Bibliographic Records       June 1995


The Information Fields

   The various fields should follow the format described below.

   <M> means Mandatory; a record without it is invalid.
   <O> means Optional.
```

The tags (aka Field-IDs) are shown in upper case.

<M> BIB-VERSION of this bibliographic records format
<M> ID
<M> ENTRY date
<O> ORGANIZATION
<O> TITLE
<O> TYPE
<O> REVISION
<O> WITHDRAW
<O> AUTHOR
<O> CORP-AUTHOR
<O> CONTACT for the author(s)
<O> DATE of publication
<O> PAGES count
<O> COPYRIGHT, permissions and disclaimers
<O> HANDLE
<O> OTHER_ACCESS
<O> RETRIEVAL
<O> KEYWORD
<O> CR-CATEGORY
<O> PERIOD
<O> SERIES
<O> MONITORING organization(s)
<O> FUNDING organization(s)
<O> CONTRACT number(s)
<O> GRANT number(s)
<O> LANGUAGE name
<O> NOTES
<O> ABSTRACT
<M> END