

RESEARCH PRIORITIES FOR THE WORLD-WIDE WEB

**Report of the NSF Workshop Sponsored by the
Information, Robotics, and Intelligent Systems Division
Held October 31, 1994 in Arlington, VA**

**This report is available on-line via:
<URL:<http://www.cc.gatech.edu/gvu/nsf-ws/report/Report.html>>**

Prepared by
**Robert C. Berwick
John M. Carroll
Chris Connolly
Jim Foley
Edward A. Fox
Tomasz Imielinski
V. S. Subrahmanian**

Edited By
Jim Foley & James Pitkow

**National Science Foundation
Arlington, VA 22230**

**The opinions expressed in this report are those of the workshop panel and
do not necessarily represent NSF policy.**

Table of Contents

Executive Summary	5
<i>Jim Foley</i>	
1. Introduction	9
<i>Jim Foley</i>	
2. Strategic Recommendations	12
<i>Robert C. Berwick, John M. Carroll, Chris Connolly, Jim Foley, Edward A. Fox, Joseph Hardin, Tomasz Imielinski, & V. S. Subrahmanian</i>	
Recommendation 1: Establish NSF WWW Office	12
Recommendation 2: Supplement PI Research in WWW Context	12
Recommendation 3: Include WWW Impact in Peer Review Criteria	13
Recommendation 4: Make WWW Standard for NSF Information Dissemination	13
Recommendation 5: Make NSF/IRIS Projects Web Accessible	14
Recommendation 6: Develop Software Repository Tools	14
Recommendation 7: Establish Software Repository	14
Recommendation 8: Establish SBIR Program	14
Recommendation 9: Integrate Research in Cognate Fields	14
3. Short Term Recommendations	15
<i>Edward A. Fox & Chris Connolly</i>	
Recommendation 1: Establish a Home Page For Each NSF Grant	15
Recommendation 2: NSF Division and Program Home Pages	15
Recommendation 3: Focused Experimental Use of the Web to Collaborate	16
Recommendation 4: Support On-line Project Proposals and Reports	16
4. Recommendation for Collaboration Tools	17
<i>John M. Carroll & Chris Connolly</i>	
Recommendation 1: Development of Link Databases	17
Recommendation 2: Reply-To and Annotation Capabilities	17
Recommendation 3: Establish Interactive Collaboration Tools	17
Recommendation 4: Requirements for Web Collaboration Tools	18
5. Research Agenda Recommendations	
5.1 Research Agenda: Indexing, Semantics & Representation	24
<i>Robert C. Berwick & Edward A. Fox</i>	
Recommendation 1: From HTML to HTIML	21
Recommendation 2: Syntax and Semantics	21
Recommendation 3: Information Brokers	22
Recommendation 4: Intelligent Agents	22
Recommendation 5: Knowledge Representation & Interchange Schemes	23
Recommendation 6: Federal Agency Efforts	23

5.2 Research Agenda: Architecture & Performance	24
<i>Edward A. Fox & Tomasz Imielinski</i>	
Recommendation 1: Improvement Techniques for Web Performance	24
Recommendation 2: Comparative Analysis of Link Specification and Maintenance	24
Recommendation 3: Comparative Analysis of Multimedia Routing Protocols	25
Recommendation 4: Modeling and Simulation of the WWW	25
5.3 Research Agenda: Resource Discovery	27
<i>Edward A. Fox, V. S. Subrahmanian & John M. Carroll</i>	
Recommendation 1: Information Retrieval Research	27
Recommendation 2: Continued NSF Support for Digital Libraries	28
Recommendation 3: Perform Needs Analysis of NSF	29
Recommendation 4: Targeted Research Areas	29
5.4 Research Agenda: Visualization	31
<i>Jim Foley & Chris Connolly</i>	
Recommendation 1: Research on Information Road Maps	31
Recommendation 2: Visualization of Bibliographic Search Results	31
Recommendation 3: Web Viewers as the New Desktop Metaphor	32
Recommendation 4: Scaling the Visualization to Different Platforms	32
Recommendation 5: Tools to Develop Context-Sensitive Visualizations	34
5.5 Research Agenda: Authoring Tools	35
<i>Jim Foley & Tomasz Imielinski</i>	
Recommendation 1: Knowledgeable Authoring Tools	35
Recommendation 2: Active Papers and their Authoring Tools	35
5.6 Research Agenda: Usability Evaluation	37
<i>John M. Carroll</i>	
Recommendation 1: Task and Information Needs Analyses	37
Recommendation 2: WWW Usability Evaluation & Methodology Research	37
Recommendation 3: Create a WWW Information “Model Farm”	38
Recommendation 4: Track the Development of the Web	38
5.7 Research Agenda: Learning & Self Organization	39
<i>Tomasz Imielinski & Robert C. Berwick</i>	
Recommendation 1: Request Reorganization and Learning	39
Recommendation 2: Service Reorganization and Learning	40
Recommendation 3: Adaptive Links	40
Recommendation 4: Different Presentation Modalities	41
6. Appendix A - Participants	42
7. Appendix B - URLs	43
Collected URLs from this Report	43
WWW Starting Points	43

Executive Summary

The World-Wide Web (WWW or Web) [Berners-Lee 92] and its associated viewers has given more substance and credibility to the Information Superhighway concept than has any other technological development since the Internet itself. It offers new opportunities for collaboration among researchers, for dissemination of information, and as an object of research. Accordingly, the Information, Robotics, and Intelligent Systems Division (IRIS) of the National Science Foundation (NSF) commissioned a workshop to provide a set of recommendations concerning:

1. Opportunities for NSF's use of the WWW for information delivery to the public and research communities.
2. Research which NSF in general and IRIS in particular should consider undertaking with respect to the Web, its accessibility, and its usability.
3. Use of the WWW as an experimental platform for collaborative efforts in the IRIS and computer science research communities, including potential enhancements to the Web in support of such collaborations.

The workshop was held on October 31, 1994, in Arlington, and was attended by six IRIS-supported PIs, five NSF staff members from IRIS, and five NSF staff members from other divisions or directorates along with three other participants - a complete attendance list is in Appendix A. This report summarizes a set of strategic recommendations for NSF and elaborates a recommended research agenda surrounding the Web.

Before discussing specific recommendations, a bit more context is appropriate. We believe that the importance of hypermedia in general and of the WWW in particular will be ranked by historians of technology in the same class as microprocessor technology, the Internet, and graphical user interfaces. Each has had or is having substantial impact on the conduct of science, education, and business; on our country's economic competitiveness; and on our society and culture.

We see the WWW as marking the start of a new era, just as did these earlier technologies. The new era will be characterized by the democratization of computer-based information and by a global, easily-used information infrastructure which will ultimately be as ubiquitous and easy to use as the dial telephone.

Indeed, the analogy to dial phones is compelling. In the era of manual switchboards, placing a call was difficult, relatively expensive, and labor intensive. As dial phones became ubiquitous, anyone could make phone calls. So too, the WWW replaces earlier, more primitive means for accessing information - FTP, Gopher, freewais, Usenet, etc. - all of which required more sophistication and knowledge to use. WWW and its browsers dramatically decreases the effort needed to access information via networks, and are changing the ways in which people use their computers, as well as the reasons they use them.

This is reflected in the dramatic growth of Web servers (from 193 in June 1993 to 1,265 in June 1994 to 13,516 in January 1995 [net.Genesis 95]), Web users (unknowable, but estimated to be in

excess of 8 million), and Web NSFNET byte traffic (from 0.002% in January 1993 to 2.6% in January 1994 to 17.6% in January 1995 [Merit 95]).

While the above numbers represent dramatic and quantitative metrics, the determination of the Web's impact on society and day to day computing has yet to be thoroughly researched. Interestingly, one of the earliest issue raised during the workshop centered around the need for societal impact and usage studies. What has made the Web so popular?

Despite the rapid growth, the Web is in a primitive state, just as early operating systems and computer networks and graphical user interfaces and microprocessors were primitive by today's standards. The Web presents many usage and research challenges, and we urge IRIS, NSF, and other government agencies to step up to the opportunities thereby presented.

Key Strategic Recommendations: We believe the most important steps for NSF to take are:

1. Establish and fund an NSF office to focus special attention on research issues concerning the WWW, and to help fund the implementation of selected Web enhancements which would greatly benefit the conduct of research, just as some years ago a special office was formed to handle networking.
2. Target supplemental funding to PIs to do their research in the context of the WWW. For instance, a program which simulates a new chip fabrication process could be written to be usable via the WWW. An innovative search engine or visualization tool could be ported to the Web.
3. Include criteria, when relevant, in peer reviews with respect to impact on the information infrastructure of science and engineering.
4. Begin an active, phased program wherein the WWW becomes the standard electronic means for NSF to disseminate information to its research communities and to the public.

Additional, but lower priority Strategic Recommendations: We also recommend that NSF seriously consider whether to:

5. Provide funding for NSF researchers to make current or past interactive systems accessible via the Web.
6. Develop software tools for researchers, educators, and librarians to use in creating and maintaining information and software repositories.
7. Fund the establishment and operation of archives of information and software of importance to specific NSF research communities, using the software tools developed under Recommendation 6.
8. Establish an SBIR (Small Business Innovative Research) program focus on the Web, as a way to further jump-start economic development surrounding the Web.

In addition, there is an obvious need for IRIS to continue to coordinate with relevant research activities in other NSF directorates, to coordinate with other agencies, and to coordinate with the inter-agency Digital Libraries Initiative [Fox 93].

Research Agenda Summary: We recommend that at least the following topics be considered for inclusion in an NSF research program focused on issues of relevance and importance to the WWW. The research agenda is elaborated in greater length in Section 5 of the report itself.

Indexing, Semantics & Representation - From HTML to HTIML; Syntax and Semantics; Information Brokers; Intelligent Agents; Federal Agency Efforts; Knowledge Representation & Interchange Schemes

Architecture & Performance - Improvement Techniques for Web Performance; Comparative Analysis of Link Specification and Maintenance; Comparative Analysis of Multimedia Routing Protocols; Modeling and Simulation of the WWW

Resource Discovery - Information Retrieval Research; Continued NSF Support for Digital Libraries; Perform Needs Analysis of NSF; Targeted Research Areas

Visualization - Research on Information Road Maps; Visualization of Bibliographic Search Results; Web Viewers as the New Desktop Metaphor; Scaling the Visualization to Different Platforms; Tools to Develop Context-Sensitive Visualizations

Authoring Tools - Knowledgeable Authoring Tools; Active Papers and their Authoring Tools

Usability Evaluation - Task and Information Needs Analyses; WWW Usability Evaluation & Methodology Research; Create a WWW Information "Model Farm"; Track the Development of the Web

Learning & Self Organization - Request Organization and Learning; Service Reorganization and Learning; Adaptive Links; Different Presentation Modalities

We would expect some of the research developments to spread into regular use on the Web, via enhancements to the Web infrastructure supported in some cases by NSF or in other cases by the private sector. Just as with Internet development, NSF will have to address the question of how much infrastructure development to support.

Why does any of this matter? One can argue that none of these research issues are really WWW issues, that they are more general and apply in one way or another to networks or to multimedia or to single-user computer systems. Such an argument is at one level correct, but at another level grossly misses the importance and impact of the Web. The Web is a widely-accessible and therefore ideally placed testbed for the issues discussed here. Doing research in the context of the Web provides a rich and unmatched environment for experimentation by allowing wider use of experimental visualizations, authoring tools, search engines, browsers, etc. The Web possesses a certain dynamic, a rapid forward movement, a context of rapid exchange of ideas which can be exploited to speed up the pace of idea generation, idea development, experimentation, dissemination, and transfer into practice. We believe that an aggressive research program surrounding the Web will be good for science, for NSF, and for the further strengthening of economic activity surrounding the Web.

A relatively modest investment by NSF in this area, if technically sound and well managed, could lead to 100-fold benefits in technology transfer, thus increasing our competitive position in world commerce, that is largely driven by capability in the information technology area. NSF should also see significant improvement in its operations (e.g. proposal handling), more positive perception by the public and Congress, and leveraging of its impact on the nation's educational system.

We thank those who helped with the workshop. As NSF considers our report, we stand ready to discuss, clarify, and further develop our recommendations.

An on-line version of this report is accessible via: <URL:<http://www.cc.gatech.edu/gvu/nsf-ws/report/Report.html>>. Note that the on-line version is supplemented with the internal and external position papers collected prior to the workshop.

References

Berners-Lee T., Cailliau R., Groff, J.-F. (1992) The World-Wide Web. *Computer Networks and ISDN Systems*, 25, North-Holland, pp.454-459.

Fox E., ed., (1993) Sourcebook on Digital Libraries: Report for the National Science Foundation. TR-93-35, VPI&SU Computer Science Dept., Blacksburg, VA. Available by anonymous FTP from directory /pub/DigitalLibrary on fox.cs.vt.edu.

Merit Network Information Center Services (1995) <URL:<ftp://nic.merit.edu/statistics/nsfnet/>>.

net.Gensis Inc. (1995) <URL:<http://www.netgen.com/cgi/comprehensive>>.

1. Introduction

Jim Foley

1995 marks the 50th anniversary of the visionary statement by President Roosevelt's science advisor, Vannevar Bush, in which he called for dealing with the information explosion through a desktop system, memex, using hypertext methods to access and add to world knowledge. His ideas and efforts to turn science into an active force for peaceful growth should inspire NSF's current plans to help build a National Information Infrastructure [Bush 45].

A fundamental and permanent change is occurring in the way computers are used. It is a change which has profound implications for our society, for our economy, for our computer industry, for our educational processes, for our work styles, for our life styles, and for our research agendas. We refer to the change from computer as stand-alone tool for word-processing, computation, data base management, to computer as gateway to the world of information and to the world of computer-mediated interaction with others.

It is a change whose seeds were sown by the email and file transfer capabilities of ARPANET, by the graphical user interfaces from Xerox PARC and their successful commercialization by Apple and then Microsoft, and by the economies afforded by microprocessor chips and modern telecommunications networks.

The World-Wide Web is the most visible and most important aspect of this change. It is the first large-scale realization of Bush's memex. As a technology, it has swept across the world like a wind-swept fire across the dry prairie. In less than two years since the introduction of the first widely-available GUI viewer (NCSA's first alpha release of X Mosaic was February 93), the WWW is currently the second largest byte and packet mover on NSFNET [Merit 95] with over 13,000 WWW servers [net.Gensis 95] accounting for over 1.8 million URLs and an estimated 10 Gigabytes of stored/generated data [Lycos 95]. We know of no comparable growth phenomenon in the computer and communications domain. Figures One and Two display the explosive growth of the WWW.

The success of the Web is clearly and directly the consequence of strategic research investments made by government agencies. The two most significant components of the success are the Internet, started initially by ARPA and sustained and nurtured by NSF, and the Mosaic viewer, developed at NCSA, the super-computer center funded by NSF. In the last year, several companies have successfully commercialized various Web browsers, creating a thriving and rapidly growing industry.

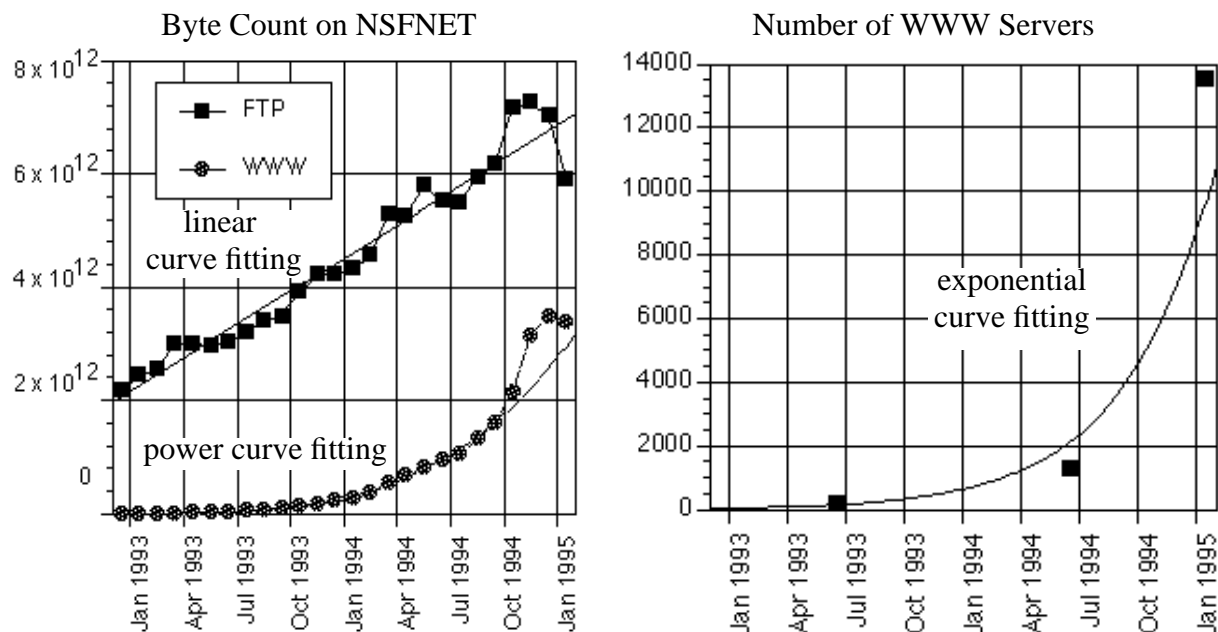
Yet the Web is in a primitive state, just as early operating systems, computer networks, graphical user interfaces, and microprocessors were primitive by today's standards. Recognizing that the Web presents many usage and research challenges, the Information, Robotics, and Intelligent Systems (IRIS) Division of the National Science Foundation commissioned a workshop to provide a set of recommendations concerning:

1. Opportunities for NSF's use of the WWW for information delivery to the public and research communities.
2. Use of the WWW as an experimental platform for collaborative efforts in the IRIS and computer science research communities, including potential enhancements to the WWW in support of such collaborations.
3. Research which NSF in general and IRIS in particular should consider undertaking with respect to the WWW, its accessibility, and its usability.

The initial workshop concept was developed within NSF by Y.T. Chen, Oscar Garcia and John Hestenes. Workshop grant IRI-9423739 was awarded to Georgia Institute of Technology to conduct the workshop and support travel and miscellaneous costs.

In preparation for the workshop, the participants prepared position papers outlining various issues surrounding the Web. These position papers are included as Appendix D of the on-line version of this report <URL:<http://www.cc.gatech.edu/gvu/nsf-ws/report/Report.html>>. Also in preparation for the workshop, the workshop description (Appendix C of the on-line version) was circulated to researchers and users, using the WWW based electronic discussion mailing lists. Responses were requested as a way to gather additional input to the workshop process. Eight replies were received; they are found in Appendix E of the on-line version of this report.

The workshop was held on October 31, 1994, in Arlington, and was attended by six IRIS-supported PIs, five NSF staff members from IRIS, and five NSF staff members from other divisions



Figures One & Two: Left, the number of bytes transferred across NSFNET with a linear fit for FTP and a power (not exponential) fit for the WWW. As of January 1995, FTP and the WWW were the first and second largest byte movers respectively. Late 1994/early 1995 drop offs are due to the changing of NSFNET architecture. Right, the number of WWW servers against an exponential curve fitting. [Merit 95 (left); net.Genesis 95 (right)]

or directorates along with three other participants. A complete attendance list is in Appendix A. This report summarizes a set of strategic recommendations for NSF and elaborates a recommended research agenda surrounding the Web.

This report is divided into four major sections:

- Strategic Recommendations for NSF action (Section 2). These are a ranked set of recommended strategies. The highest-priority strategic recommendation is that NSF establish and fund an office to focus special attention on research issues concerning the WWW, and to help fund the implementation of selected Web enhancements which would greatly benefit the conduct of research, just as some years ago a special office was formed to handle networking.
- Short-term recommendations (Section 3). These are immediate internal actions which NSF might take to enhance use of the Web by PIs, by NSF, and as a means of disseminating NSF information to the public.
- Recommendations for tool-building opportunities (Section 4). Good tools, if available, would increase collaboration over the Web by NSF PIs and the scientific community in general. We recommend that NSF consider funding the development and deployment of some or all of these tools.
- A research agenda designed to further develop the science and engineering basis on which the Web is built (Section 5). We recommend that, whenever feasible, the research actually be conducted on and with the Web, so that it can benefit from immediate interaction with other Web researchers and users. The basic idea is to build a strong community of Web-oriented researchers.

We thank all of the NSF staff members who were involved in the workshop, colleagues who responded to our requests for ideas and comments, and our individual support staffs and NSF support staff who facilitated our travel and participation in the workshop. Ms. Joan Morton and Ms. Chrissy Whitaker of Georgia Tech's GVV Center administered the workshop grant.

As NSF considers our report, we stand ready to discuss, clarify and further develop our recommendations.

References

Bush V. (1945) As We May Think. *Atlantic Monthly*, 176, pp.101-108.

LycosTM <URL:<http://lycos.cs.cmu.edu/>>

Merit Network Information Center Services (1995) <URL:<ftp://nic.merit.edu/statistics/nsfnet/>>.

net.Gensis Inc. (1995) <URL:<http://www.netgen.com/cgi/comprehensive>>.

2. Strategic Recommendations

**Robert C. Berwick, John M. Carroll, Chris Connolly, Jim Foley,
Edward A. Fox, Joseph Hardin, Tomasz Imielinski, & V. S. Subrahmanian**

During the latter part of the workshop, the invited PI participants (see Appendix A) enumerated a set of strategies to recommend to NSF as ways to take advantage of the opportunities presented by the Web. They then ranked the strategies by relative importance; the order in which the strategic recommendations are enumerated indicates their ranking, with the first being the most important.

Key Strategic Recommendations: We believe the most important steps for NSF to make are:

Strategic Recommendation 1: Establish NSF WWW Office

Establish and fund an NSF office to focus special attention on research issues concerning the WWW, and to help fund the implementation of selected Web enhancements which would greatly benefit the conduct of research, just as some years ago a special office was formed to handle networking. We believe the Web is of sufficient importance that this needs to be at the level of an office as opposed to a single program manager. This is the very strongest and highest priority recommendation of the workshop participants.

The most basic reason we recommend a focal-point office within NSF is the opportunity NSF has for further enhancing the impact of the Web. This impact has several dimensions. First, the impact on the conduct of general scientific research can only be enhanced, not diminished, by developing collaboration tools which work with and integrate into the Web. Second, an extensive computer science research agenda, as elaborated in Section 5, surrounds the Web. Much of this research can in some ways be done on its own, without using the Web. But, as part of the next recommendation, we argue that doing research using the Web as an extended laboratory is good science. Third, the economic development and international competitiveness aspects of increasingly sophisticated Web capabilities, driven by research with the Web, is not to be taken lightly.

Creating a focused Web office can help ensure an intentional research agenda which serves these purposes. The research agenda described in other parts of this report should fully or partially be the responsibility of this office.

Strategic Recommendation 2: Supplement PI Research in WWW Context

Target supplemental funding to PIs to do their research in the context of the WWW. For instance, a program which simulates a new chip fabrication process could be re-written to be usable via the Web. A number of the research agenda recommendations can lead to new software, such as browsers, authoring tools, and search engines - which would be of great value if widely distributed for use on the Web. A funding mechanism should be developed to support hardening of research prototypes into relatively stable (not necessarily commercial quality) distributable tools for use by a broader set of users than the immediate research community which might use a new tool on an experimental basis. In some cases, NSF should then support integration of new capabilities into non-commercial servers and browsers.

The advantage of moving research prototypes onto the Web is that experimental systems can quickly be tried out by many users. For instance, Mauldin's LycosTM search engine <URL:http://lycos.cs.cmu.edu/> receives nearly one half million requests a day as of March 7, 1995. Usage such as this can provide researchers with rapid feedback on the utility and robustness of their tools, algorithms and interfaces. Many fundamental research issues are best framed in the context of heavy usage in unpredicted ways. Understanding how real people use tools is far more revealing of their tools strengths and weaknesses than experiments in a traditional lab setting. In some sense, a Darwinian selection can occur earlier in the scientific research cycle, by exposing tools more quickly to 'the light of day' offered by more general use via the Web. But, placing tools on the Web also takes more resources than working in more traditional ways.

Consider the growth in operating system research once UNIX was developed and source code was available to the research community. UNIX provided a lab environment for experimentation. The lab environment flourished even more after Berkeley UNIX was developed (with government research funding). The Internet and its preceding ARPANET had a similar effect on networking research.

A second advantage is that the availability of research prototypes on the Web increases the rate of technology diffusion by making ideas more visible, more quickly.

This is not to say that a research concept in its most embryonic stage should be tried out on the Web. Certainly the most basic ideas can and should be developed in a more controlled environment. But, researchers should be encouraged and supported in moving their embryonic systems onto the Web.

Strategic Recommendation 3: Include WWW Impact in Peer Review Criteria

Include criteria, when relevant, in peer review of proposals with respect to impact on the information infrastructure of science and engineering. Just as current review criteria include impact on science and human resources, so too, we believe, NSF should ask reviewers to consider, where appropriate, how positively the proposed research will affect the Web.

Strategic Recommendation 4: Make WWW Standard for NSF Information Dissemination

Begin an active, phased program wherein the WWW becomes the standard electronic means for NSF to disseminate information to its research communities and to the public. This is the longer-term focus of the modest beginnings discussed in our short-term recommendations which we believe NSF should make right now (and in some cases is already beginning to make). Completing the process for all of NSF, as opposed to beginning the process for a selected subset of computer-intense NSF programs or divisions, will require considerable resources and should be the subjected to debate and scrutiny.

Additional, but lower priority Strategic Recommendations: We also recommend that NSF consider whether to:

Strategic Recommendation 5: Make NSF/IRIS Projects Web Accessible

Provide funding for IRIS researchers to make current or past interactive systems accessible via the Web. This would involve retrofitting current work to the Web. It is ranked lower than the focus of Strategy 2 above because planning for Web use with new projects is potentially more efficient than retrofitting of older programs.

Strategic Recommendation 6: Develop Software Repository Tools

Develop software tools for researchers, educators, and librarians to use in creating and maintaining information and software repositories.

Strategic Recommendation 7: Establish Software Repository

Fund the establishment and operation of archives of information and software of importance to specific NSF research communities, using the software tools developed under Recommendation 6.

Strategic Recommendation 8: Establish SBIR Program

Establish an SBIR (Small Business Innovative Research) program focus surrounding the Web, as a way to further jump-start economic development surrounding the Web.

Initially this should be supported by special supplemental awards. NSF should also begin to formulate revised guidelines for awards, and open up those revisions for discussion by the community to ensure enthusiastic cooperation by PIs, who will then take on these responsibilities as part of the standard obligations of awardees.

Strategic Recommendation 9: Integrate Research in Cognate Fields

NSF should encourage related research in cognate fields and integration of their results into the WWW by adding to the funding of IRIS's programs which deal with:

hypertext and hypermedia, information capture, storage and retrieval including multimedia, electronic publishing, human-computer interaction involving information access, distributed information systems, collaboration technology, knowledge models, intelligent agents, adapting to users - varied age/experience, multicultural, multilingual, computational linguistics.

This should be done in conjunction with revising program guidelines to include specific mention of the importance of applying research to the development of the WWW.

3. Short Term Recommendations

Edward A. Fox & Chris Connolly

In this section we discuss modest low-cost immediate actions which we believe should be taken within NSF to facilitate use of the Web. We also note that as a pre- or co-requisite to implement these recommendations, NSF will need to ensure that its own Web infrastructure is capable of supporting access by NSF personnel, by the PI community, and eventually by the general public. NSF will need to adequately and appropriately educate, equip and support its personnel for use of the Web in order to implement these recommendations.

Recommendation 1: Establish a Home Page For Each NSF Grant

NSF should develop a Foundation-wide initiative to encourage PIs to publish information about their NSF-sponsored projects. The project Home Page should have links to all the computer-based artifacts created under the grant. This would typically include published papers and book chapters, technical reports, dissertations and theses, bibliographies, software, data sets, images, videos, audio segments, animations, interactive demonstrations, simulations, talks and lectures, the proposal for the research project, and the final report for the project. The project Home Page should also link to Home Pages for the project PI(s) and GRA(s) and to the academic unit(s) of the PI(s). The process of developing these pages will need to be evolutionary, starting with modest information content but growing over time. Implementing this Recommendation will allow researchers to access up-to-date technical information from their colleagues, and will set expectations for new PIs, as well as assist prospective PIs in formulation their research proposals.

Operationalizing this Recommendation should start in the divisions whose PIs are the most computer-literate, and could be encouraged by:

- Program directors interacting with their PIs;
- Letters sent to PIs as part of the new award package;
- Developing a comprehensive set of example templates for the Home Pages and guidelines defining appropriate content;
- Conducting a competition for the “best” set of project pages.

Recommendation 2: NSF Division and Program Home Pages

Home Pages should be established for each NSF division and program. Each program’s Home Page should have links to the program description, to the Home Pages of each of the current and recent projects, workshops, centers, etc. supported by that program, the Home Page of the program director, and to the division Home Page.

Recommendation 3: Focused Experimental Use of the Web to Collaborate

NSF should pick a model Division, and one or more model areas, for experimentation regarding use of the WWW. Thus, the PIs working on some grand challenge application should be assisted and encouraged to make extensive use of the Web, and this whole process should be carefully studied (e.g., considering usability issues, developing a design history of the process, identifying social and organizational changes).

Recommendation 4: Support On-line Project Proposals and Reports

NSF proposal processing can become a model World-Wide Web management information system. Instead of moving tons of paper around the country to distribute announcements, pre-proposals, full proposals, reviews, and reports, they should be distributed, collated, and posted on the Web. This is the natural way for the current STIS system to evolve, and indeed it must be an evolutionary process, in order to deal with the implementation, infrastructure, and personnel training costs. The results, eventually, should decrease some current operational costs. Further, some aspects of this can't be fully implemented until security is incorporated into Web clients and servers - a trend which is already underway.

4. Recommendation for Collaboration Tools

John M. Carroll & Chris Connolly

In this section we discuss recommendations for tools, which if created and deployed would enable greater collaboration among NSF PIs and among scientific and other communities in general. The World-Wide Web can facilitate asynchronous and synchronous collaboration among researchers in a variety of ways. As suggested in the Short-Term Recommendations section, it allows the creation of standard PI Home Pages with links from project abstracts posted by NSF to PI Home Pages, and it enables on-line submission, review and distribution of NSF project proposals and reports. It can support the development of a world wide distributed digital library of scientific and technical information, as discussed in the strategy section, including development of link databases to facilitate querying document links. It allows the use of reply-to and public annotation capabilities for technical information posted on the Web, as well as interactive collaboration tools (e.g., a World-Wide Web Interactive Talk (WIT; see <URL:http://info.cern.ch/hypertext/WWW/WIT/User/Overview.html>) capability, and object-oriented multi-user discussion environments (MOOs)).

Several of our recommendations will require active involvement by NSF Web users. Hence, as also discussed in the section on short term recommendations, adequate infrastructure, training, and personnel support at NSF will be needed.

Recommendation 1: Development of Link Databases

An important aspect of collaboration is knowing who has linked to posted information, or what other documents point to a posted document. Link databases associated with posted documents would allow the investigator who authored a document to notify those who have linked to his or her documents of revisions or follow-on work. They would allow researchers to access the set of documents that refer to the current document. Such an architecture would make documents on the Web part of a real network of knowledge. A version of this capability is already implemented in Hyper-G [Kappe 93; Kappe 94], and should be made available on the Web (e.g., [Pitkow 95]).

Recommendation 2: Reply-To and Annotation Capabilities

Encourage development and oversee use of reply-to and public annotation capabilities for technical information posted on the Web. Investigators should be able to communicate directly concerning posted information, with one another and with sponsoring NSF program managers. For example, it should be easy to reply to a currently viewed document, by generating an e-mail message to the PI or the program manager. It should be possible to publicly annotate posted documents (though this may need to be managed by NSF). Creating such direct links would encourage direct collaboration among investigators.

Recommendation 3: Establish Interactive Collaboration Tools

Many Web facilities are being developed specifically for collaboration, both synchronous and asynchronous. For example, there are a variety of debate forums, based for the most part on WIT.

WIT uses the forms capability of HTML to support debates structured in ways similar to Rittel's classic IBIS (Issue-Based Information System [Rittel 73]). Users state contrasting positions on various issues and adduce evidence and arguments to make their cases. Two current limitations with this capability are (1) that in many cases its actual use has devolved into serving merely as a better interface for general newsgroups, and (2) that it only supports linking flat text objects. More focused WIT collaborations have been created by incorporate management schemes, for example in the Virginia Tech computer ethics debates <URL:<http://info.cs.vt.edu:8000/debates/html/Debates.html>>. Research on approaches to moderating these interactions could increase their potential utility as tools for collaboration within the scientific community. Extending the WIT framework to support linking documents and bibliographies as backing for positions taken would also help make this a more suitable tool for scientific collaboration.

The most common tools for synchronous collaboration are multi-user discussions, such as BioMOO, a MOO for collaboration among biologists <URL:<http://bioinformatics.weizmann.ac.il:70/1s/biomoo>>. BioMOO hosts informal discussions, presentations, and even journal clubs. All of the participants in BioMOO enter remotely, and at any one time are likely to be scattered throughout the globe. The result is a rapid exchange of ideas and results which heretofore would only occur at annual meetings or the occasional workshop. Currently, most MOOs are exclusively text-based and accessed via telnet. BioMOO is an example of a MOO that can exploit Web browsers, however the activities it supports that go beyond text-based interaction (for example, the 24-hour posters), are all asynchronous. An important avenue of exploration and development is the support of higher bandwidth synchronous collaboration.

A pioneering example is the USC Mercury Project, which allows users to remotely command an excavating robot <URL:<http://www.usc.edu/dept/raiders/>>. In general, the Web can be used to allow researchers to remotely access exotic equipment at other sites. For example, researchers wishing access to range data could conceivably access sites which have available range sensors, which would then capture images and transfer them to the requesting researcher. As another example, remote access to anthropomorphic hands or other robotic devices could enhance the development of algorithms for dextrous manipulation. This remote access capability should be useful in any discipline which requires data collection using exotic or expensive hardware. An important direction for development is the support of greater interactivity; collaboration in the Mercury Project is implemented via fairly cumbersome turn-taking.

Recommendation 4: Requirements for Web Collaboration Tools

The actual use of Web-based collaboration tools is not well-understood: What are the tools used for now? How effective are they? What are the routine problems that users encounter? Studies are needed of scientific collaborations such as those currently supported by BioMOO and the USC Mercury Project. These collaborative communities are on a different scale than typical computer-support cooperative work (CSCW) projects, and it is likely that their fundamental objectives and concerns are just different. Unless we focus systematic effort at understanding these communities, developments in support for scientific collaboration using the Web will be no better than evolutionary.

References

Kappe F., & Maurer, H. (1993) Hyper-G: A Large Universal Hypermedia System and some Spin-offs. *Computer Graphics* (experimental on-line issue).

Kappe F., Andrews K., Faschingbauer J., Gaisbauer M., Pichler M., & Schipflinger J. (1994) Hyper-G: A Network Tool for Distributed Hypermedia, (on-line documentation) <URL:ftp://iicm.tu-graz.ac.at/pub/Hyper-G/doc/>.

Pitkow J., & Jones R. K. (1995) Towards an Intelligent Publishing Environment, *Computer Networks and ISDN Systems*, 27, North-Holland (in press).

Rittel H., & Webber M. (1973) Dilemmas in a General Theory of Planning. *Policy Science*, 4, pp.155-169.

5.1 Indexing, Semantics & Representation

Robert C. Berwick & Edward A. Fox

Discovering resources and retrieving information (see also Section 5.3) in the WWW depends upon indexing of information to capture the underlying semantics, and representing that to facilitate search and browsing. Today, the best retrieval systems for the Web rely on programs to walk regions of the Web and build indices for statistical retrieval based on each document's full text. Such automatic meta-indexes or "indexes of indexes" like the Configurable Unified Search Engine CUSI <URL:<http://web.nexor.co.uk/public/cusi/doc/search.html>> provide a partially credible global indexing platform, but because they walk over a static snapshot of the Web at any one time, they cannot adapt to dynamic aspects of the Web, such as documents generated on-the-fly. The reasons for the current inadequate state of WWW indexing are straightforward: Because it takes a long time to walk the Web (three months as of March 1994), global indices are out of date before they are compiled -- and the time required to traverse the entire Web is increasing, presumably, exponentially, in lockstep with the growth of information on the Web. In general, Web walkers cannot index pages which are synthesized on the fly or are gatewayed from other information systems (e.g., searches in relational or text databases).

Retrieval of sounds, images, speech, and text, while highly desirable, remains a research goal. In general, weaker indexing methods are used: statistical retrieval, using manual indexing. There is currently no commonly agreed upon method (the WAIS protocol leaves open the search engine to be used), as does GILS (Government Information Locator System, OMB's system built on the Z39.50 protocol). There is ample room for improving effectiveness of all Web-based retrieval systems through better indexing. The following discussion identifies particular problem areas and then offers suitable recommendations to improve indexing on the WWW.

Quantity and dynamics. The amount of networked information is very large and growing very rapidly. This presents a serious problem with updates (see above) that has yet to be adequately solved. It also makes developing "perfect searches" (those that have very high precision) very difficult; users will be forced to develop expertise at searching or spend more of their time looking at imperfectly ranked result sets unless better indexing and search systems are developed. Faster networks and faster computers will help, but like building better highways, they will in all likelihood just increase traffic flow; system administrators have learned this lesson many times about disk space.

Redundancy. The same information often appears in many forms, and needs to be distilled down to a non-redundant account. People use the redundancy of natural language to pack multiple descriptions into a few sentences. In the WWW there is redundancy from mirror sites, multiple formats, multiple media types, and successive versions. Note that minimization of redundancy does not necessarily conflict with the notion of providing multiple views of the same information.

Indexing not in infrastructure. Although the HTTP protocol provides for typed links, these are not yet widely used to create meta-level assertions about documents. New documents are not added to site or global indexes upon entry. Additionally, the piggybacking of queries into the syntax of URLs does not scale well to large queries.

Recommendation 1: From HTML to HTIML

Just as the simplicity of HTML stimulated the development of WWW, NSF should encourage the formulation of a HyperText Indexing Markup Language (HTIML). Such an “open standard” should encourage indexing. Paragraph-based indexing engines should be supplied as part of the software available for the WWW (see Recommendation 3 below). We propose that NSF sponsor a workshop to improve indexing of the WWW, considering approaches like HTIML. The HTIML proposal is to make something like an artificial English a “standard” for index markup, because the redundancy of English (or any other natural language) makes it useful as a highly redundant way to specify queries used to retrieve text.

(a) Note that like HTML, HTIML need not be perfect; rather, it simply has to be open-ended enough to allow for appropriate extensions (including indexing of multimedia information). In the near term (next 6 months), one way to bootstrap HTIML would be simply to provide an additional templating capacity focussed on indexing.

(b) Paragraph-template indices can be generated automatically and then retrieved by full-text index queries, as in the START¹ system. Thus, users should be able to retrieve pages via queries like, “Which researcher at the AI lab works on mobile robots?” In addition, by annotating free-form text with English phrases and sentences, then matching these annotations with incoming queries, the power of sentence-level natural language processing can be effectively put to use in the service of information retrieval. Furthermore, this technique generalizes easily to the indexing and retrieval of all types of information, whether or not these are based on text.

(c) HTIML could be proposed as an international standard to help encourage human indexing of WWW documents. This approach should be considered also as an enhancement, supplement, or addition to the current suite of standards for document markup. Thus, HyTime adds architectural forms to SGML to manage hypertext and multimedia, and might be developed to work with any SGML document needing indexing. Then, HTIML could be viewed as an extension to HTML as well as a stand-alone language to help with index construction of any type of information.

Recommendation 2. Syntax and Semantics

Research should be encouraged on specifying the linkage between syntax and semantics in documents marked up with SGML, and on how to improve searching with that information. While SGML allows description of document structure, and DSSSL allows description of the mapping from tagged documents to a layout presentation, there is no description of the mapping from tagged documents to a semantic representation. A specification for such mappings would allow documents to be indexed in a fashion that would consider semantic intent. Thus, names or dates in

1. The START natural language system (SynTactic Analysis using Reversible Transformations) consists of two modules which share the same Grammar. The understanding module analyzes English text and produces a knowledge base which incorporates information found in the text. Given an appropriate segment of the knowledge base, the generating module produces English sentences. A user can retrieve the information stored in the knowledge base by querying it in English. The system will then produce an English response.

documents identified syntactically could be easily indexed by converting them to a suitable canonical representation. Later, at search time, a “query-by-example” scheme could ask for documents with a given name occurring in the list of authors, or in some bibliographic reference. Eventually, this scheme could become a standard for semantic specification to supplement the SGML suite of standards.

Recommendation 3: Information Brokers

Research should be encouraged to fit indexing, link data, and later retrieval into the distributed environment of the WWW. Traditionally, retrieval systems work with large databases and index all the data they contain. In the WWW this usually means that a server will do this for information it contains, building an index, and providing a search system to access that index and its data. Alternatively, some “worm” may collect a large amount of WWW information, index that, and provide search support for it, along with pointers to the original location. However, there are other scenarios that should be explored. For example, the Harvest approach allows for collection of data from a group of servers, indexing of that data, and searching in parallel against a number of such group servers or information brokers. Similarly, the Whois++ and Uniform Resource Characteristics (URC) approach by the Internet Engineering Task Force enables both global resource location and discovery via a hierarchical infrastructure of index passing and result caching. Other options deal with: where and how data is collected, where and how indexes are stored, and what degree of parallelism and centralization is involved in searching. Further, it is likely that some data will be indexed in several ways: e.g., for different purposes, perhaps with different depth or detail or processing involved. Thus, some systems might index a numeric table as a string of values while another might index it as row-label/column-label/value triples.

Given the rich link structure of the Web, that data should be reflected in the indexing of documents. Thus, links into and going out from a document characterize it, and should be used as part of the indexing, as is done with citation search systems, or those that include indexing of citation and bibliographic coupling data along with term indexing.

Research should consider how these various collection, selection, analysis, indexing, and retrieval methods can be integrated. Toolkits are needed to help, as are protocols, that would fit into a general scalable architectural framework. This must be related to natural language analysis methods and query interfaces.

Recommendation 4: Intelligent Agents

Indexing and interface methods should be extended to better support intelligent agents. Research is needed to bring together work on intelligent agents with other work on indexing and document analysis, to lead to new architectures, approaches, representations, and methods. Right now there is little interchange between workers adopting these different approaches. There are separate problems addressed by each group, such as handling knowledge representation and protocol for agent, or handling large index collections. However, considering these problems together can have great synergy. Thus, the Z39.50 standard for information retrieval in client-server situations could be extended to support richer document characterization and easier processing by agents. Many agents look for specific patterns or structures in data, and may use what they find to help

with further analysis; this approach could work well with regular search systems handling very large collections.

On the interface side, agents should be developed that have a rich interaction with users. In some cases, indexing can assist, if linguistic structures (e.g., phrases) or syntactic elements (e.g, title, author list) are known to be required during searching. Thus, if an agent works with a user to disambiguate a word, identifying the precise word sense desired, that can be of greatest value if indexing was carried out to the sense level. Usability studies (see Section 5.6) are needed to tie in with this research.

Recommendation 5: Knowledge Representation & Interchange Schemes

We recommend research into: knowledge representation and interchange schemes in a distributed environment of autonomous agents; upgrading of speech act theory into a theory of collaboration in large “invisible colleges”; applying knowledge bases to tailor and increase the efficiency of the lifelong learning experience of large numbers of individuals; developing “research environments” (e.g., for a chemist, or a numerical analyst) that integrate a variety of tools with related documentation as well as datasets and information resources; and constructing diagnostic systems that help humans or computers try to determine causes and repairs for various common types of failures (e.g., of equipment, of social groups, of economic enterprises). All of these studies and more will be enabled by the accumulation of vast information stores into a comprehensive information architecture. This type of basic research is needed to enable their efficient utilization by millions of future users of a World-Wide Knowledge Web, with reasonable performance. All of these investigations must occur in the environment of an open network, with methods and standards evolving as required regarding: representation, interchange, protocols, and interfaces. Interoperability is the key, and the WWW provides an environment inherited from the Internet where rapid prototyping and large-scale testing of small inventions is encouraged.

Recommendation 6: Federal Agency Efforts

NSF should work proactively to support other Federal Government efforts that involve or should require more effective indexing and searching, starting with running a workshop on this topic. Many government agencies manage large collections of information, and engage in indexing activities. In the case of the National Library of Medicine, this involves extensive manual indexing as well as a number of automated indexer aids. In the case of DoD, there are the TIPSTER/TREC and MUC activities that include testbed development as well as studies of indexing, message understanding, document analysis, ad hoc retrieval, and routing. For the White House, there are efforts to manage electronic mail. For the Intelligence and Defense Communities there are numerous efforts for filtering, routing and retrieval. NSF should encourage collaboration by running a cross agency workshop to consider distributed indexing issues, to handle very large collections, semantic searching, intelligent agents, and other approaches in the context of the WWW, but informed by other studies and investigations such as those mentioned above, or new efforts with Digital Libraries.

5.2 Architecture & Performance

Edward A. Fox & Tomasz Imielinski

As the Internet community rapidly grows, usage of WWW continues to increase, and more multimedia information becomes available, it is essential that innovative research and development projects focus on the resulting architectural and performance problems that will threaten the viability of the Web. Already we see: servers swamped when some information they provide is suddenly “discovered”; long delays or timeouts when accessing popular Web pages; intolerably slow transmission of images, audio and video files; frequent error messages indicating a link has been “broken” (e.g. by moving the target file); and wasteful re-transmission of large files from remote sites, even when requested almost immediately after a prior access is complete.

Some of the existing problems can be easily solved through training, especially if organizations’ “Webmasters” have proper guidelines. Thus, making use of proxy servers can reduce re-transmission and traffic through caching, video files can be split into a number of smaller segments to reduce startup delays, images can be compressed using the JPEG standard to cut down on transmission time, and systems with link databases (e.g., Hyper-G [Kappe 94] & WWW based [Pitkow 95]) can ensure consistency of link and node sets. Clearly, education and training is needed.

Recommendation 1: Improvement Techniques for Web Performance

We recommend research, development and training efforts to help discover and validate techniques to improve Web performance, e.g., determination of optimal configurations of servers and proxy servers, along with settings of parameters (e.g., cache size, garbage collection strategies) on both client and server systems, for organization-wide use of WWW technology; prediction of performance that takes into account the very skewed access distributions commonly associated with hypertext collections; preparation of authoring guidelines regarding in-line inclusion, sizes of pages, segmentation of multimedia streams, and summaries or lower resolution versions of large objects; and development of courseware and other training materials to disseminate these results (see also Recommendation 4 below).

Problems like: broken links; inadequate semantics and searchability of links; and impossibility of attaching source anchors to read-only items can be resolved if link databases become part of the basic infrastructure of the WWW, instead of a subparts of it (e.g., the collection of Hyper-G systems).

Recommendation 2: Comparative Analysis of Link Specification and Maintenance

We recommend research and comparative studies regarding the current WWW scheme for specifying links, vs. alternative schemes such as having a distributed link database. These investigations relate to the fundamental architecture of the WWW, its performance, its support of searching methods (e.g., over link attributes and metadata), and user interface functions like browsing the node-link graph.

While some Internet services use broadcast methods (e.g., listserv), the WWW adopts a global hypermedia model where objects are delivered on demand (after supplying some universal identifier). For real-time video distribution over the Internet (e.g., MBONE), multicast methods are applied. Hence, we must consider: What changes will be needed as WWW usage shifts more to the use of multimedia information, where objects are very large, are constituted of streams that require synchronization, and in some cases must satisfy real-time delivery demands in order to provide adequate quality of service? How will other Internet services change to be integrated into the Web as it evolves to become the “memory” for the networks?

Recommendation 3: Comparative Analysis of Multimedia Routing Protocols

Therefore, we recommend research and comparative studies regarding WWW and its connection with: multicast; real-time information delivery; on-demand services; large multimedia objects and streams requiring synchronization; listservs and other types of asynchronous or synchronous conferences.

These investigations will help ensure that the WWW evolves to better support the growing demands for computer tele-conferences and other synchronous communications, as well as more common electronic mail, computer conference, bulletin board, listserv, and USENET services - all enhanced to include multimedia information. This line of research will be essential to determine: what user demands and requirements will be for such services; what new workloads are likely to arise; and how various architectures, algorithms and protocols influence performance and quality of service.

All of the above mentioned problems take on new dimensions as the WWW scales up, in terms of amount of information, size of information items, number of authors, number of computers, number of users, and amount of usage by each person. It is this continuing pressure on existing infrastructure that must be anticipated so that research studies will yield new ways to improve the WWW architecture and its performance. Some should deal with networking concerns, others with algorithms, and yet others with database systems, data structures, file systems, information systems, or knowledge bases. Other studies should consider whether and where indexes are kept, what analyses are pre-computed vs. prepared by knowbots, what search heuristics are developed and deployed, and where replication schemes are applied.

Recommendation 4: Modeling and Simulation of the WWW

The recommendations above (especially Recommendation 1) indicate the need for research regarding modeling and simulation of the WWW, including developing workloads (based on logs and traffic measurements), describing architectures, exploring the implications of information organization, and comparing approaches to information processing and management. Devices like caching, prefetching, clustering, classification, and filtering should be considered to determine their effectiveness in reducing traffic and improving service.

When these results are obtained and then refined through empirical studies, it will be feasible to begin to address more complex problems. In particular, as the WWW transitions from a data to an information and thence to a knowledge network, with ever more powerful systems providing

desired services, it will be necessary to consider the interaction of agents and other intelligent modules. In-depth content analysis (e.g., using natural language or image processing techniques), summarization, planning, constraint satisfaction, translation and conversion, and other highly complex tasks will need to be supported on a grand scale.

All of these investigations must occur in the environment of an open network, with standards evolving as required regarding: representation, interchange, protocols, and interfaces. Interoperability is the key, and the WWW provides an environment inherited from the Internet where rapid prototyping and large-scale testing of small inventions is encouraged.

References

Kappe F., Andrews K., Faschingbauer J., Gaisbauer M., Pichler M., & Schipflinger J. (1994) Hyper-G: A Network Tool for Distributed Hypermedia. (on-line documentation) <URL:ftp://iicm.tu-graz.ac.at/pub/Hyper-G/doc/>.

Pitkow J., & Jones R. K. (1995) Towards an Intelligent Publishing Environment. *Computer Networks and ISDN Systems*, 27, North-Holland (in press).

5.3 Resource Discovery

Edward A. Fox, V. S. Subrahmanian & John M. Carroll

Finding the right information in the Web is a difficult problem, especially in light of the “information explosion” and rapid growth of the Web. Similarly, finding the right information in a Digital Library [Fox 94; Fox 95] is also a hard problem. Even finding the right information in a large collection is difficult. Though in one sense these three tasks are all embodiments of the generic problem of “information retrieval”, they differ in terms of scale, supporting methodologies, availability of supplemental or meta information, complexities related to the requirement for distributed processing, and performance demands. In the WWW, resource discovery builds upon what was discussed in Section 5.1 regarding Indexing, Semantics & Representation.

For the foreseeable future, working on the problem of information retrieval in large collections will be of value. There are continual improvements in this field, and allied areas (e.g., natural language processing) also continue to yield new tools and techniques that can be tested and integrated with other approaches. Connecting with hypertext, browsing, and visualization are all important parts of the solution. Success here is certainly of value in solving similar problems for the WWW.

Recommendation 1: Information Retrieval Research

We recommend ongoing support for research into “information retrieval” and the fields of hypertext, multimedia systems, and human-computer interaction, especially as they relate to the problem of finding information in large collections. Advances in this regard usually can be applied directly to the larger problems of finding information in digital libraries or the WWW, since it seems that many advanced retrieval methods have scaled up after only minor refinement (see experiences in connection with the ARPA funded TIPSTER and TREC projects). Yet, new problems also arise in regard to the next level situation, that of digital libraries.

Digital libraries are large, and usually are made up of a number of heterogeneous collections. They will typically include at least a moderate amount of multimedia information, which presents new challenges relative to only searching text. They make it more important to manage variety in item size. In addition to indexing, cataloging is required, usually to handle large items or item collections. Other problems come to roost, like eliminating duplicates or relating multiple versions.

Since traditional libraries are often located and planned to serve a range of user needs, such as university education and research, digital libraries will certainly be required to support important user tasks. Part of this means facilitating search, navigation and browsing in and between multiple spaces, such as document space, concept space, keyword space, and others relating to strings, features, or linguistic constructs. Whether one deals with geographic information, video archives, or data from outer space, many of the same general approaches will apply, especially if integrated with domain specific processing.

Adapting these general approaches to particular types of collections, however, warrants empirical validation in a variety of representative situations. It is also necessary to develop a theory and

practice of specification of complex collections and their processing, whereby: specialized methods can be applied to each multimedia data type; new approaches can be developed to handle the complex interrelationships of those data items into common structures (e.g., multimedia streams and performances, composite documents, multimodal interaction schema); and distributed retrieval operations can effectively operate upon such large and complex heterogeneous collections (as found in digital libraries).

Digital libraries are possible now in part due to advances in electronic publishing (interpreted in the broad sense). People create most documents with computers, and also capture a large percentage of multimedia objects into digital forms. With proper standards, much of what is needed by way of indexing, cataloging and metadata can be automatically recorded in usable forms. Such standards and practices can be applied to the WWW at large, especially if urgently needed research develops tools and techniques that allow domain-specialized standards to be used in general purpose environments (like the Web) without losing their expressive power or requiring labor-intensive coding or translations.

With so much automatic processing, the trend to disintermediation will continue. Thus, an important challenge in developing digital libraries is that of preserving as much as possible of the knowledge and skills expert intermediaries demonstrate. Library and information science have much to contribute, and knowledge acquisition in this regard is of particular value. In addition, models, architectures, prototypes and tests regarding agents must be undertaken to see if such knowledge can be properly applied. Hence:

Recommendation 2: Continued NSF Support for Digital Libraries

We recommend ongoing support by NSF for research on digital libraries. Just like the ARPA TIPSTER project's utility was greatly enhanced by involving many smaller project groups through the TREC efforts, a significant amount of support should be provided for investigators not funded under the Digital Library Initiative, since they are likely to make important contributions, using the new testbeds that are already under development. It would be helpful if these new projects could build upon the DLI testbeds, accessing part of the corpora, using software libraries, or analyzing sanitized access logs. Important leveraging will result if new investigators can visit the DLI sites in person or electronically, participate as colleagues in special workshops, or directly make use of the developing DLI research facilities.

Another way to think about the future WWW collection of NSF-related information is as a digital library for science and engineering research and education. As such, it deserves serious study, and requires careful coordination and attention. There are enumerable important resources that will be included, and which will support user tasks related to education and research. Program managers have particular information needs¹: managing their own information and also accessing information about the scientific community to find reviewers, panelists, and related activities of other agencies, as well as to contact investigators and task force attendees. The resource discovery needs of investigators operate in reverse, since they must learn about NSF programs, program managers, previous awards, funded proposals, project reports, and a myriad of related documents.

1. See position statement by J. Hestenes in Appendix D of the on-line version of this report.

They may find it helpful to locate and communicate with investigators funded by NSF, or others who have attended NSF-sponsored workshops. Clearly, it is possible to describe many of the needs of program managers or investigators. Much harder is the task of determining the resource discovery needs of the rest of the nation in terms of NSF-related information. Hence:

Recommendation 3: Perform Needs Analysis of NSF

We recommend that NSF arrange for a careful analysis of the data, information and knowledge needs of its staff, of investigators in the sciences and engineering, and of others in the nation who might benefit from accessing what should evolve into a digital library of NSF-related materials. This resource discovery needs analysis should deal not only with content but also with organization, task specification, access and manipulation capabilities, and user interface characteristics.

With this analysis complete, NSF will have a roadmap to build, through efforts of those it employs or funds, a significant digital library. As this evolves, it will relate to larger and larger portions of the WWW. Thus, it is important for NSF's operations that it encourages the improvement of resource discovery methods on the WWW. It is also of benefit for other uses of the WWW for NSF to provide support. In particular, there is now little funded research dealing with the many scientific problems tied to resource discovery on the WWW. Hence:

Recommendation 4: Targeted Research Areas

We recommend that NSF fund research aimed at discovering useful resources in the WWW:

- Methods of widely distributed search, and subsequent fusion of results;
- Problems of parallel search on the scale of the Web and adapting gracefully to the many types of failures that can occur on such a large and complex networked system;
- Techniques to find good starting points or to identify particularly promising collections where more focused searching can begin;
- Approaches to computer-assisted browsing, that helps users quickly narrow the scope of their investigation;
- Architectures for organizing information and processing elements, such as those using agents, or those working with separate layers (e.g., gatherers, brokers, harvesters, centroids);
- Models for searching in such contexts that are marked by extremes of heterogeneity, wide variety of partial structures, and diverse genre;
- Integrating the link database of the WWW with results of node analysis and indexing for improved search, navigation, visualization and browsing, allowing for wide variations of node in-degree and out-degree, and inconsistencies, as well as both typed and untyped links;

- Constructing limited and isolated distributed test environments, for benchmark comparison purposes, “protecting” the rest of the WWW from high-risk experimental studies, and to help with validation of the abovementioned research (especially parallel search and simulation studies); and
- Developing educational and training materials regarding use of advanced information retrieval and resource discovery methods and tools.

Some joint workshops involving researchers dealing with document models and searching, object technologies, and various types of scientific databases might stimulate work on the above areas in the context of important applications (e.g., genomics, social and behavioral database research, and digital libraries).

References

Fox E., Akscyn R., Furuta R., & Leggett J. (1995) Guest Editors' Introduction to Digital Libraries. *Communications of the ACM*, 38(4), (in press).

Fox E., ed., (1993) Sourcebook on Digital Libraries: Report for the National Science Foundation. TR-93-35, VPI&SU Computer Science Dept., Blacksburg, VA. Available by anonymous FTP from directory /pub/DigitalLibrary on fox.cs.vt.edu.

5.4 Visualization

Jim Foley & Chris Connolly

Information visualization, like its cousin scientific data visualization, is a powerful way to effectively convey understanding to the user. Recognizing that the WWW is an enormous graph of nodes and arcs, we imagine various overview and navigational views of WWW sub-structures being presented to the user to assist in exploratory resource discovery. For instance, a graph showing the relationship of a home page to all of its subsidiary pages would provide a user with a valuable “road map” to all of the pages. However, the paucity of explicit content semantics (meta-information about nodes and links) in HTML makes creating meaningful overview maps quite difficult.

Recommendation 1: Research on Information Road Maps

Support research on ways to visualize WWW-like information spaces. To do this, at least three issues need to be addressed: 1) What types of “road maps” are useful for navigating? 2) What descriptive semantics must be available in HTML or related structures (link databases, structures above HTML, etc.) to allow the road maps to contain meaningful semantics? 3) How can the road maps be automatically generated from the semantics?

A reasonable place to start this work is with graph-like diagrams, but graphs are not the only visual metaphor for information space navigation. The way-finding literature in architecture, map-making, book design, and information design all offer rich sets of concepts. Now is the time to explore such metaphors to understand their applicability. Those which show promise should be tried, both in formal experiments and experientially. Means to automatically generate an appropriate metaphor for a particular information space will be needed: such generation includes both creation of the visual representation itself as well as the automatic binding of individual information elements to the visual metaphor.

Because information spaces are so large, an important element of creating overviews will be creating appropriate visual abstractions which convey meaningful relationships while eliding unnecessary detail. This in turn will often require semantic descriptors beyond what is found in HTML or in other meta languages.

Recommendation 2: Visualization of Bibliographic Search Results

Support research to develop and experiment with a wide variety of ways to present and manipulate search results. Presenting the user with a linearly-ordered list of documents based on some notion of “relevance” is quite abstract and does not help the user understand the high-dimensional space within which relevance is defined. Various scatter plot and other types of presentations have been developed [Heath 95; Rao 95; Verasamy 95]; exploratory research is needed to develop others, and experimental studies are sorely needed to understand the strengths and weaknesses of the existing visualizations. A set of benchmark tasks against which to measure various visualizations is another needed research outcome.

Most bibliographic search facilities currently used on the Web rely on string search rather than more sophisticated database organization and query techniques. Integration of more advanced information retrieval tools into the Web would enhance bibliography management and sharing. The set of current visualization tools for retrievals should be integrated into WWW search engines, as a way of exposing a wider community to the state of the art and hence engaging their imaginations in creating new tools.

Recommendation 3: Web Viewers as the New Desktop Metaphor

Another aspect of visualization deals with the potential for expanding the role of Web viewers. Web viewers present to a computer user a collection of resources which are available on other people's computers - currently, on other Web servers. At the same time, the user has to look, in other windows, at a different collection of different visual representations of other types of resources: programs, files, directories, email messages, etc. located on his or her own computer. This asymmetry is paradoxical. After all, one of the key Web concepts is the URL, the Universal Resource Locator. Why should URLs not be used to locate resources in a truly universal manner, including those on one's own workstation? Why should a Web viewer not present the total set of resources available to a user? Hyper-G provides a local file browser [Kappe 93].

The idea, then is to use the Web and its viewers as a single, unified mechanism for presenting and accessing information, eliminating the artificial separation between one's own local files which are not integrated into a Web server, and universal resources, which are integrated.

This has several implications, some of which deal with visualization and presentation, others of which go deeper within the Web infrastructure. At the presentation level, it reinforces the need for a research agenda dealing with visual metaphors for presenting, understanding, organizing, and navigating rich information spaces. Note that visual interfaces to computer operating systems, such as the Macintosh OS and Microsoft Windows, provide simple but inadequate such metaphors. There is no reason to think that other metaphors are not available. Global file systems, like Prospero [Neuman 92], which provide simplistic browsing, are indeed research in the right direction.

Within the Web infrastructure, the "Visualization as universal resource locator" concept means that programs have to become first-class Web objects, objects which can be searched for and loaded and executed, just as text files can be searched for and loaded. In this way the Web becomes a browsable repository for programs, opening up local file systems to authorized users via a single, powerful mechanism. Incorporation of programs as objects would be a useful adjunct to the active paper concept, since each program is a complete algorithm specification. This greatly facilitates the reproduction and sharing of research results in applied disciplines such as robotics and computer vision.

Recommendation 4: Scaling the Visualization to Different Platforms

Access to the Web is currently from standard computing platforms, e.g. UNIX workstations, PCs, Macs, etc. With the increasing use of PDAs, PC-based TV set-top boxes, and mobile computers communicating via limited bandwidth, research is needed to develop methods of automatically

scaling, or adapting, Web information to these new platforms. Also, as more and more individuals with handicaps access information via computers, the same scaling concepts can be used to adapt presentations to assist those with limited sight or limited motor skills. We discuss here three aspects to scalability: display real-estate scaling, bandwidth scaling, and auditory scaling.

Real estate scaling is concerned with the limited screen resolution of PDAs and home TVs. The research agenda here includes developing rule-based information presentation tools whose task is to automatically render information in ways which make sense for the targeted device - by eliding detail, by creating hierarchical levels of selection where none previously existed, by spatially rearranging information, by whatever it takes to allow the user to sensibly access and see (or hear) information.

Bandwidth scaling refers to being able to deal with limited bandwidth between the information source (server) and information destination (client). This is an issue even for workstations connected at gigabit or ethernet speeds, and becomes even more of a concern for a 19.2Kbps wireless connection or a 9.6Kbps home telephone connection (we believe the demise of this low-speed access will be much slower than predicted). One approach to bandwidth scaling is multiresolution retrieval and display, whereby information displayed to the user is incrementally refined as more data is retrieved and components are received at progressively higher resolutions. Each information object published by a server may be defined at various resolutions according to its data type. For a digitized photograph, a low resolution version may have several adjacent pixels grouped together and assigned an average color value (e.g., an image pyramid). For a large Postscript document, a low resolution copy may consist of the abstract and the index. Depending on the typical load on the server and the storage requirements, an agent at the server site can decide to store one or more of the multiresolution versions and generate the others on demand. In the default mode of object transfer, a low resolution copy of the object is retrieved and displayed to the user while the full or higher resolution copy is being transferred. The requesting user is thus shown a version of the object promptly and can decide whether to continue with the retrieval or to abort the request. This can noticeably reduce the 'dead time' spent by users as they wait for network connections to be made and the requested information to be transferred. Similar retrieval and caching schemes are already used in some image analysis systems, but can be generalized in the Web to apply to many different object types.

Audio scaling is the concept of either replacing or supplementing visual information with auditory cues (or vice versa), or adapting audio to monaural, stereophonic, quadrasonic outputs and to different acoustic environments.

In the coming days of ubiquitous and mobile computing, as an individual frequently moves from use of one platform to another, all types of scaling used will need to be consistent and provide seamless integration across platforms.

Tools to automate the process of building views will help make the Web more scalable with respect to the rate of generation of new data and bandwidth. Agents may be deployed at the user's platform to scan the information that has newly become available from a set of given sources and selectively incorporate it into the user's personal view according to a specified criterion.

The navigational views of information space will need to be integrated with this notion of multi-resolution retrieval as well. In a very real sense, a visual map of an information space is simply a low resolution representation of the space. Intelligent agents will be needed to create all sorts of abstracted views of individual information documents and of documents organized in an information space.

Recommendation 5: Tools to Develop Context-Sensitive Visualizations

Database systems have long applied the concept of subschemas to provide multiple views of the same underlying information. One way to think of multiple views is in the context of the scalability issues just discussed. But another way to think of multiple views is in the context of access-path view dependency. For example, the GVU Center at Georgia Tech has an extensive set of pages with a distinctive GVU look <URL: <http://www.cc.gatech.edu/gvu/>>. However, the home departments of GVU faculty and students all have home pages with their own looks. It is unreasonable to ask everyone to have multiple sets of pages, one with the GVU look and one for their home department's look, with the pages having duplicate information but having different graphic design looks and different logical organizations of information onto pages and different linkage structures. What is needed is a way to separate content from appearance, organization, and linkages so that the access path to an information collection can be used to determine which "view" to apply to the underlying information. Note that this notion of views is also relevant to the discussion of automatic reorganization in Section 5.7.

References

- Heath L., Hix D., Nowell L., Wake W., Averboch G., & Fox E. (1995) Envision: A User-Centered Database from the Computer Science Literature. *Communications of the ACM*, 38(4), (in press).
- Kappe F., Maurer H., & Scherbakov N. (1993) Hyper-G: A Universal Hypermedia System. *Journal of Educational Multimedia and Hypermedia*, 2(1) pp.39-66.
- Neuman B. C. (1992) Prospero: A Tool for Organizing Internet Resources. *Electronic Networking: Research, Applications and Policy*, 2(1), pp.30-37.
- Rao R., Pedersen J.O., Mackinlay J.D., Masinter L., Hearst M., Halvorsen P.-K., & Card S.K. (1995) Towards Rich Interaction in the Digital Library. *Communications of the ACM*, 38(4), (in press).
- Veerasamy A., Hudson S., & Navathe S. (1995) Visual Interface for Textual Information Retrieval System. *IFIP 2.6 Working Conference on Visual Database Systems - 3*.

5.5 Authoring Tools

Jim Foley & Tomasz Imielinski

Authoring (creating) material to be presented in an information space involves at least three inter-related, overlapping, and non-sequential activities:

- Develop the overall logical structure of the information. This can be envisioned as creating a graph in which nodes are pages (documents) and arcs are links.
- Define the content of each page or document.
- Design the visual appearance (layout) of each page.

Basic authoring tools are becoming available: they tend to concentrate on layout, with little assistance for overall logical structuring. They generally take the reasonable approach of using existing graphic and text editors for developing content. None of the current or envisioned tools actively help the author, nor do they facilitate or encourage innovation.

Recommendation 1: Knowledgeable Authoring Tools

Support research to develop authoring tools which actively help and encourage the author to create good material. Just as word-processors support writing with spell-checkers and grammar checkers, so should authoring tools help with the authoring task. But, because Web pages are (or can be) more highly structured than text, the support for authoring can be at a much more sophisticated level than spell and grammar checking.

One example of this is in the domain of education. Much is known about how to structure effective educational presentations: using examples to illustrate or to motivate principles; using counter-examples to explain categorizations or an exception to a rule, etc. This leads to the notion of “pedagogically-informed authoring tools” for creating Web content, wherein good educational practices are embedded in the tools in such a way as to make it easy to do things pedagogically well, and to make it hard to do things pedagogically poorly. This philosophy is already used in some experimental user interface design tools which can critique a design, [Foley 89] and should be extended to Web authoring tools.

As with the visualization research issues, having a rich semantic description is as well an essential ingredient for this work.

Recommendation 2: Active Papers and their Authoring Tools

One concept for educational and information-sharing use of the Web is that of active papers, papers which can be visualized, rendered, explained and tested. Active papers are not static but rather are snapshots of the evolving experimental research which allow the “readers” to follow the footsteps of the authors by recreating the reported experiments and even modifying them. Active papers can use visualization and speech and can be queried by the interface which is specific to

the paper content. Therefore, the active paper is a time- evolving process and reading such a paper involves not only querying its current content but also monitoring its future. Active papers provide the readers with the possibility of interacting with the content of the paper using the interface provided by the author (in the same way as she now provides the table of contents). In this way, each paper becomes a “database” which can be queried and even, in some cases, updated. As in databases, different views can be defined for the same paper: For example one can create “presentation views” for talks of different length and depth including high level (possibly visual) abstracts of “what the paper is about?”

The active paper will be only a snapshot of the ongoing research process. Active papers will be extensible both by authors (producing new versions) and by other scientists by producing active hyperlinks. Thus reading the active paper will involve also monitoring the “follow up” changes made to it. An issue closely related to this is the refereeing of such active papers. NSF and professional societies can be instrumental in this by establishing central repositories or databases of pointers (URLs) to active papers. Admission of active papers to such repositories would be contingent on periodic peer review.

References

Foley J., Kim W., Kovacevic S., & Murray K. (1989) The User Interface Design Environment - A Computer Aided Software Engineering Tool for the User-Computer Interface. *IEEE Software, Special Issue on User Interface Software*, 6(1), pp.25-32.

5.6 Usability Evaluation

John M. Carroll

The WWW offers many possibilities for the support of computer science research and the dissemination of technical and administrative information by NSF. Few of these possibilities are technologically routine, and all would entail significant change in the established practices of computer scientists and NSF personnel. Thus, it is critical that the NSF develop an explicit plan to monitor and analyze the usability of various elements of WWW technology as they are deployed.

Recommendation 1: Task and Information Needs Analyses

We recommend support of the development of example task-analyses and information-needs analyses for PI homepages. A basic element proposed in this report is to encourage all PIs to develop appropriately designed and linked homepages. However, currently we do not have an explicit understanding of the characteristic tasks and needs of the users of WWW homepages: What are people trying to do, how are they trying to do it? What are the problems? How do these tasks, needs, and problems vary as a function of the type of user (a computer scientist, an NSF Program Manager, a newspaper reporter)? Of course, such analyses are coupled to the state-of-the-art: the use of and requirements for WWW information will have to be tracked dynamically as innovations in WWW technology and practices are introduced.

Recommendation 2: WWW Usability Evaluation & Methodology Research

The most pertinent task-analyses and information-needs analyses will, of course, be empirically grounded. Today, many appeals to “use” of the Web refer merely to large access logs. It is important to know that Web pages are accessed, of course, but they can be accessed for many different reasons - including misleading links and navigational errors. We need to promote a detailed understanding of what people are doing with the Web, how they are using it, and what their real needs and problems are.

For NSF’s immediate concerns, some subset of PI homepages should be instrumented to create a database of usage for subsequent analysis and modelling. This, in turn, raises many research questions regarding methods and instrumentation for remote evaluation, and policy issues regarding informed consent. An important example of remote evaluation methodology is Georgia Tech’s Graphic’s Visualization & Usability Center’s extensive research with surveys <URL:http://www.cc.gatech.edu/gvu/user_surveys/User_Survey_Home.html>. Another example is Virginia Tech’s adaptation of the critical incident method <URL:<http://hci.ise.vt.edu/cgi-bin/wwwproj/story>>.

The result of this effort should be an explicit set of Web page requirements. These requirements should be exemplified in prototype page designs: this will ensure that the requirements are operational and facilitate their application in the short-term, and will allow the requirements themselves to be directly evaluated empirically in the longer-term.

Recommendation 3: Create a WWW Information “Model Farm”

Understanding and facilitating new uses of the WWW can be integrated through a focus on highly visible “model farm” projects that would be extensively instrumented, closely documented, and fully accessible to the World-Wide Web community. Such projects would provide an beacon both to the PI community and to the public at large with respect to NSF’s efforts at modernizing its information dissemination. They would convey concrete models of the process of managing and utilizing Web information resources. The best candidates would be NSF projects that involve substantial project management, for example, through the collaboration of several institutions or groups -- the six new digital library projects are excellent candidates for this.

Recommendation 4: Track the Development of the Web

In the Introduction, Foley refers to the magnitude of potential change unleashed by the Web. It is essential for us to keep reminding ourselves that these changes may in fact be historic - the technologies we use every day become mundane to us very quickly. The institutions of science are clearly transforming themselves, and this transformation may in the end be quite profound. At the same time, because this transformation pertains to communication and collaboration across electronic networks, the tokens and patterns of the change are easily captured and organized. A systematic effort should be undertaken to build a history of the World-Wide Web with respect to its role in the management and conduct of science.

5.7 Learning & Self Organization

Tomasz Imielinski & Robert C. Berwick

It is critical that the WWW's hyperlink-based information structure dynamically learn usage patterns so that:

1. Frequently requested information becomes easily accessible: high-demand links are cached or pushed to the top, requiring few link traversals, possibly at the expense of rarely-demanded information.
2. Services migrate where they are most needed: servers dynamically replicate close to the areas of highest request. This calls for dynamic replication, migration, and caching schemes. Using self-adaptation, the WWW will be better able to automatically tailor itself to individual user needs, as well as handle a larger traffic volume.

Note that the WWW "structure" is understood here in two different senses: (1) as a graph with links as edges, where links themselves can be dynamically defined in response to the usage patterns; and (2) as a mapping between services and servers that provides these graph structure. Hence a service can dynamically migrate to a server as result of a growing request volume. Both scenarios of "dynamic restructuring" can lead to optimization: optimization of expected number of link traversals in order to reach a page and optimization of network traffic and overall latency (response time per page).

We recommend that NSF support progressive evolution to an "adaptive" WWW in two ways:

First, by adopting when possible available adaptive/learning add-on technologies to web page accessors/servers (see below); Second, by initiating tests of the adaptive approach, in conjunction with user modeling efforts. We now address both kinds of self-organization - request and service based - in more detail.

Recommendation 1: Request Reorganization and Learning

In the same way that Huffman coding assigns the shortest codes to the most frequently used symbols, a good hyperlink based interface should assign "shortest paths" to the most frequently accessed pages. Dynamic reorganization schemes are needed that use known request-per-page statistical distributions to reorganize the hyperlink interface; these algorithms can minimize the expected number of link traversals per user's query.

Techniques used for building "good" or suboptimal classification trees such as C4.5 can be used here. Individual links can "die" if there is very little or no usage. Similarly, "macrolinks" can be created and even automatically replace frequently traversed paths. In this way the WWW server can "evolve" with changing access patterns with or without human intervention. A "good" server should characterize itself by a low expected access path length.

The client's interface should also reflect the particular client's profile. It is important that the

user's list of preferred pages or "hotlist" is not just a flat collection but is itself a hyperlink-based structure. But how to provide a hierarchy on the top of a flat collection of pages? This is similar to a classification problem or even to the clustering problem in unsupervised learning. It would be very helpful if "suggested" classification trees could be provided to the user, who then would choose the one that suits them best. Note that some existing Home Pages already include user-interactive classification-tree algorithms that could be easily adapted to test this idea. We therefore suggest that NSF possibly adopt one of these methods perhaps in a separate experimental "adaptive" Home Page, that researchers could test. A larger-scale experiment with such a system is also possible. Creating such "user's-eye" WWW views as a kind of user's "miniweb" is an important research area, related to the key problem of good user modeling.

Recommendation 2: Service Reorganization and Learning

Research in dynamic replication, migration and caching can be utilized in deciding how WWW servers should replicate their services. The best schemes would allow service replication closest to high service demand areas. Pages could migrate between different servers in response to the changing demand patterns. Typically, global increase in a page's read activity will result in bringing this page closer to the center of "reading activity". If writing and updating is allowed as well, it will constitute a "counterforce" against a growing number of replicas (the update cost is higher). Dynamic replication methods has been developed and studied for data allocation in multiprocessor architectures - some of these same techniques can be directly utilized in the WWW setting.

Internet is a unique environment - there is no other distributed system of such scale, that at the same time has already accumulated such substantial usage statistics. It is critical to use these data dynamically in order to adaptively configure services to minimize access time and network traffic.

Furthermore, the WWW's rich information resources are currently accessible only to a relatively small subgroup, yet all should have equal access. The reasons for current limitations are straightforward: WWW's current interaction mode demands a computer terminal with sufficient available bandwidth, hence reasonable response time. Broadening the Web's constituency demands a wider range of user interfaces. Technologically, this will be surely possible, since in the future we should be able to support a much wider spectrum of client devices, including battery powered palmtops with narrowband wireless connection, and units with voice transcription or even vibration-mode output. In fact, the future WWW client need not be a computer user at all, but merely someone using a fax, phone, or television.

In order to support a varying spectrum of information modalities, we have to extend authoring tools and technologies such as HTML to define different page presentation styles, much as "font styles" and "style sheets" are defined today. In the best case, such an extended modality authoring system could automatically adapt to the end user's preferences.

Recommendation 3: Adaptive Links

Since Web users may vary from broadband/fixed network-based to mobile and wireless connected, we need to provide different views of the same information depending on the "resources

of the user". For example, a JPEG image of a page may take too much time and energy for a mobile client to download. Perhaps a grayscale outline (or "greeked" representation, as publishers call it) or even a textual description of the same page would be more appropriate. In general, a HTML link should be parameterized by the client's resources such as bandwidth of the connection, energy source (steady or battery power), buffer size, and the like. The actual link which is involved will be dynamically determined by the server on the basis of current "amount" of client's resources. This will put heavier burdens on information service designers, who will have to provide alternative "views" of the same information. Each page's presentation will have attached "resource prerequisites" that are necessary for satisfactory access.

It is a simple extension of this idea to provide for modality-challenged individuals. As a simple first step, modality-tagged preferences can be flagged for hearing/sight-impaired users as part of their default preferences; this ranges from simple extensions such as automatically encoding volume/font size preferences to a public-access "reader" utilities comparable to current commercial text-to-speech systems. Not all cross-modality conversions are equally simple; note that being able to "describe" full images via natural language is an area of current research, unlikely to be generally possible in the near-term. This means that for the present, off-line tagging of image-based text is likely to remain in the domain of HTML authoring and document meta-information generation. We suggest that NSF explore this range of possibilities via an augmentation similar to TDD facilities that are supplied in other contexts.

Recommendation 4: Different Presentation Modalities

We argue that frequently requested information pages should be periodically multicasted rather than provided "on demand." For example, the airline schedule for flights departing within the next hour should be periodically broadcasted to wireless users in a particular airport area due to the anticipated high demand for such information. The same information will be provided only "on demand" further from the airport. This periodic multicast may be called a "publishing" mode as opposed to the Web's currently standard "on demand" client-server paradigm. Note that "publishing" mode does not require uplink transmissions from the client. Thus, it is the only possible way to access information for users with "one way" communication abilities such as "one way" pagers. For instance a popular service such as Motorola's Embarc periodically broadcasts information such as financial news to users equipped with PDA's (HP 100LX) and extended with special one way pagers.

In the wireless infrastructure, different cells, managed by so called Mobile Support Stations (MSS), will have autonomy regarding the presentation of each page of information. The same page may be published in one cell and provided on demand in a different cell. This will require an ability to "handoff" from the client who will cross the border between two cells, when their "hot list" of pages may have a different delivery mode.

The ability to specify the mode of page presentation should be a part of a future authoring environment. The author of such a page (or perhaps its "local distributor" in case we deal with the wireless environment) should also have an option to specify other parameters. For example, if a page is to be periodically published, how often should it be published? Under what conditions? Similarly, modality conversions can be specified.

Appendix A - Participants

<u>Person</u>	<u>Organization</u>	<u>E-mail</u>	<u>Phone</u>
Robert Berwick	MIT	berwick@ai.mit.edu	617 - 253 - 8918
Larry Brandt	NSF/CISE/DASC	lbrandt@nsf.gov	703 - 306 - 1963
Su-Shing Chen	NSF/ITO	schen@nsf.gov	703 - 306 - 1927
John Carroll	Virginia Tech	carroll@cs.vt.edu	703 - 231 - 6931
Chris Connolly	U Mass - Amherst	connolly@psyche.mit.edu	617 - 253 - 5740
Jim Foley	GVU/Georgia Tech	foley@cc.gatech.edu	404 - 853 - 0671
Edward Fox	Virginia Tech	fox@vt.edu	703 - 231 - 5113
Oscar Garcia	NSF/IRIS	ogarcia@nsf.gov	703 - 306 - 1922
David Garver	NSF/DIS	dgarver@nsf.gov	703 - 306 - 1160
Jeff Graber	NSF/EHR	jgraber@nsf.gov	703 - 308 - 1165
Steve Griffin	NSF/IRIS	sgriffin@nsf.gov	703 - 306 - 1930
Joseph Hardin	NCSA	hardin@ncsa.uiuc.edu	217 - 244 - 7802
John Hestenes	NSF/IRIS	jhastene@nsf.gov	703 - 306 - 1930
Tomasz Imielinski	Rutgers	imielinski@cs.rutgers.edu	908 - 445 - 3551
Hans Joseph	CRCG	hjoseph@crcg.edu	401 - 453 - 6363
Howard Moraff	NSF/IRIS	hmoraff@nsf.gov	703 - 231 - 6931
Jim Pitkow	GVU/Georgia Tech	pitkow@cc.gatech.edu	
Larry Reeker	NSF/IRIS	lreeker@nsf.gov	703 - 306 - 1962
Maria Zemankova	NSF/DBES	mzemanko@nsf.gov	703 - 306 - 1926

Appendix B - URLs

Collected below are the URL's referenced throughout this report. These have been collected to facilitate browsing and exploration. Afterwards, pointers to WWW starting points for beginners and software are offered. This report is available on-line <URL:<http://www.cc.gatech.edu/gvu/nsf-ws/report/Report.html>>

Collected URLs from this Report

BioMoo <URL:<http://bioinformatics.weizmann.ac.il:70/1s/biomoo>>

Configurable Unified Search Engine (CUSI) <URL:<http://web.nexor.co.uk/public/cusi/doc/search.html>>

GVU Center Home Page <URL:<http://www.cc.gatech.edu/gvu/>>

GVU'S WWW User Surveys <URL:http://www.cc.gatech.edu/gvu/user_surveys/User_Survey_Home.html>

Lycos™ <URL:<http://lycos.cs.cmu.edu/>>

Virginia Tech CS Debates <URL:<http://info.cs.vt.edu:8000/debates/html/Debates.html>>

Virginia Tech Critical Incident Method <URL:<http://hci.ise.vt.edu/cgi-bin/wwwproj/story>>

USC Mercury Project <URL:<http://www.usc.edu/dept/raiders/>>

WWW Interactive Talk <URL:<http://info.cern.ch/hypertext/WWW/WIT/User/Overview.html>>

WWW Starting Points

Entering the World-Wide Web: A Guide to Cyberspace - An Introductory Guide to the W3 <<ftp://ftp.eit.com/pub/web/guide/>>

HTML Beginner's Guide <URL:<http://www.ncsa.uiuc.edu/General/Internet/WWW/HTML-Primer.html>>

NCSA's Mosaic Home Page <URL:<http://www.ncsa.uiuc.edu/SDG/Software/Mosaic/NCSAMosaicHome.html>>

WWW Frequently Asked Questions (FAQ) <URL:http://sunsite.unc.edu/boutell/faq/www_faq.html>

WWW Home Page <URL:<http://www.w3.org/>>

WWW Software <URL:<http://info.cern.ch/hypertext/WWW/Status.html>>