

VIRTUAL WORKING GROUPS AT NCEAS: USING THE WEB TO FACILITATE SCIENTIFIC COLLABORATION

Mark P. Schildhauer

National Center for Ecological Analysis and Synthesis, 735 State St., Suite 300,
Santa Barbara, California, 93101

Abstract. Advances in computing and networking are creating new ways for scientists to engage in long-distance collaborations. This chapter describes how the National Center for Ecological Analysis and Synthesis (NCEAS) is using the World Wide Web and other network services to enable *ad hoc* teams of ecologists to share ideas and data. It also discusses how technological developments in the near-future will further increase ecologists' abilities to interact using the Internet.

INTRODUCTION

Ecologists are rapidly becoming aware of the advantages of using internetwork technologies in the service of their science. In less than ten years, email has risen from being an exotic high-end application to being almost universally adopted for routine communications among scientists. Within the last five years, the World Wide Web (WWW, web) has become an extremely popular and convenient 'place' to present scientific resources, including data, preprints, and references. Currently, there is intense interest in using the Internet for easy querying and access of ecological data. Ecologists can further benefit by broadening their perspective on internetwork technologies, especially regarding the possibilities for doing effective collaborative science.

Software that supports collaboration is categorized as *groupware* or *CSCW*-- for 'computer-supported cooperative work'. CSCW researchers are addressing issues that facilitate computer-based collaboration, including location independence, synchronous (real-time) and asynchronous communication, coordination within and among groups, and information workflow. CSCW is a rapidly developing area, and many of the specific solutions discussed in this paper will be outmoded within several years by more robust, easy-to-use, and comprehensive alternatives. Nevertheless, there are already immediate advantages that can be attained by adopting groupware approaches. Specific groupware solutions currently in use at NCEAS are the focus of this chapter. These are largely workaround solutions that involve imaginative deployment of standard services. The intent of this chapter is to familiarize ecologists with the potential and basic concepts of groupware, as well as the technological underpinnings for these services. The tone is non-technical, so that the contents will be understandable to any ecologist with only minimal background and interest in Internet technologies.

PROBLEM

A primary mission of NCEAS (<http://www.nceas.ucsb.edu>) is to facilitate integrative and synthetic research in the ecological and environmental sciences. NCEAS supports collaboration among groups of scientists with complementary and cross-disciplinary expertise. NCEAS-sponsored groups may be comprised of a few to several dozen individuals, who are associated with a diversity of institutions. These scientists might not normally engage in close collaborative efforts. Since NCEAS does not support the collection of new raw data, these ventures also frequently require collection and collation of relevant information from many sources. These newly

aggregated data then serve as the basis for analyses that will hopefully lead to novel, integrative insights.

A challenge to the NCEAS' computing team was how to enable dozens of *ad hoc* working groups of scientists to effectively collaborate--prior to, during, and following any workshops or conferences at our physical facilities in Santa Barbara, California. We had to consider that our potential clients represent the entire national (and in many cases, international) community of ecological research scientists. These individuals have access to highly varying levels of computing, technical support, and network bandwidth at their home institutions. We also had to accommodate the need for scientists to work comfortably with familiar technological tools at their home sites, while providing a centralized service linking them together as a collaborative effort through NCEAS.

SOLUTIONS, 1996-1997

Development of a 'shared virtual working environment' was approached through deployment of readily available software, in conjunction with special configurations of computing servers. For example, these services allow anyone with a standard WWW browser, coupled with access to email and ftp, to participate in a collaborative networked research environment, that allows for efficient dissemination, updating, and browsing of data files, routine communications, analytical approaches, formatted works-in-progress, and supportive graphics. Before describing these mechanisms in greater detail, we provide some background as to our design requirements, and approaches to solving these issues for a large, diverse, distributed user base.

Prerequisites: network bandwidth and computational power

We started with what was feasible for the end-user, given the current level of Internet access and computer expertise of research ecologists circa January 1996, when NCEAS started hosting groups of collaborating scientists. We had to make assumptions as to computational power as well as the network bandwidth available to a typical individual in the academic or governmental sectors, since most of our clients are research ecologists working at universities, field stations and research laboratories, and federal or state agencies.

It was assumed that individuals have reasonable connectivity with the Internet—meaning that most have 24-hour access from their research offices, with faster than modem-speed (approx. 28.8 kilobits/sec) bandwidth. In fact, we were aware that a few individuals would only have intermittent modem-level access--especially those working in remote field sites, but the great majority of scientists coming through NCEAS are from institutions with at least T1-level connections to the Internet (Table 1).

Table 1. Suitability of different network bandwidths for varying types of applications, assuming that networks are not congested. The relative speed of ‘1’ for the modem is based on the currently popular 28.8kbps speed. Each successive row encompasses all the functionality of lower speed access types.

Access type	Speed	Task suitability	Typical location
Modem	1	email, simple WWW	Home phone
T1	54	graphical WWW, small data transfer	Remote office
Ethernet	347	Limited videoconferencing, shared file systems	Existing LAN
Fast Ethernet	3,470	streaming video, large data access	Emerging LAN
Gigabit Ethernet	34,700	multi-channel multimedia, live data feed	LAN in 3-5 years
ATM/SONET: OC3-OC48	5,382- 86,111	Integrated voice/video/data; guaranteed Quality of Service (QoS)	Next generation Internet; LAN in 5-10 years?

Our problem is complicated by the dispersed distribution of our clientele. This means that although individuals might have good connectivity within their local area network (LAN), severe bottlenecks can occur anywhere between their desktop and the ultimate destination of NCEAS’ servers. High-speed access to the Internet, however, is a necessary prerequisite for advancing groupware approaches among *distributed* (remotely located) individuals. Luckily, current interest in the Internet among the US federal government (<http://www.ngi.gov>), academicians, and the commercial sector, enables us to forecast that next generation internetworking initiatives, currently exemplified by ‘experimental’ networks such as Internet-2 (<http://www.internet2.edu>) or the vBNS (<http://www.vbns.net>), will provide vastly increased bandwidth to most academically-based ecological researchers within the next several years (Table 1).

We also had to make assumptions about the level of computational power routinely available to NCEAS’ clients at their home institutions. It was decided that most individuals were familiar with email software and a graphics-based WWW browser (e.g., Netscape® as opposed to Lynx), and that this was an increasing trend. Given this supposition, we developed solutions requiring minimal configurations roughly equivalent to an Intel 486/DX2-66 running Windows 3.1, or PowerMac at 66MHZ running MacOS System 7 or higher (Table 2). The machines would need a minimum of 8MB of RAM, and ideally 16MB or more. These machines would also need a TCP/IP networking stack installed on them, to work through an Ethernet card or via SLIP or PPP through a modem.

Our minimal system configuration represents technology that would have been common on a scientist’s desktop circa 1994, i.e., about two years prior to our launching of NCEAS’ collaborative work areas. We expect a large part of our client base has since upgraded their desktops to more powerful systems.

Table 2. Computational power available to typical ecological researcher, using PC's as an example, and assuming cost of approximately \$2500 at time of purchase. Intel power comparisons are estimated from iComp 2.0 measurements (<http://www.ideasinternational.com/benchmark/bench.html>).

Computer	CPU and RAM	Relative Power	'Purchase' Date
PC	Intel 486- 66 MHz, 8MB	1	Early 1994
PC	Intel Pentium- 133 MHz, 16MB	3.7	Early 1996
PC	Intel Pentium II- 300 MHz, 64MB	11.2	Early 1998

Client/Server model

The client/server model forms the basis for all currently popular Internet services, and was the model we turned to for providing services to NCEAS' distributed user-base. *Servers* answer the 'requests' for service from a potentially large number of client systems at any given time. If these are critical services, the more powerful hardware at a site will be allocated for these purposes, with the expectation of accurate and reliable service backed by a trained professional staff. Server software usually requires considerable expertise in order to be properly configured. This unfortunately prevents most individual ecologists from setting up systems that would enable them to easily 'publish' on the WWW, or even facilitate simple file transfer among remote colleagues without compromising their own system's security. *Client* systems, on the other hand, run software that is relatively easy to install and learn, but only useful if capable of connecting to a server. So, e.g., while many scientists are finding that it is quite simple to upgrade and configure their Web browser, they are still unable to implement some very effective server-side functions unless provided by local systems administrators. In essence, the client/server model enables an organization to leverage the capabilities of a few high-powered, secure, and well-tuned servers to deliver information to a broad user base.

Any computer can function as a server if the appropriate software is installed. But on today's Internet, most of the large, powerful servers are UNIX systems, with multiple CPUs, true multitasking and multi-user capabilities, and highly optimized throughput. The main services provided by today's Internet servers include email, such as the SMTP-based 'sendmail' included with virtually every UNIX system; Web serving via programs such as Apache or Netscape FastTrack™; and less obvious but critical functions such as the Internet Domain Name Service (DNS), which enables clients to use computer names (hostnames) rather than numerical IP addresses to locate other machines on the Internet.

The rapid and continuing success of the Internet over the past decade arises as a consequence of the still growing number of inter-connected servers running standard services for local groups of users (e.g., at departmental or campus levels), while at the same time easy-to-use clients for these services enable networked desktop systems to run useful 'proto-groupware' such as email or the WWW. It is also an indication of the infancy of the Internet's mainstream acceptance that servers are still relatively difficult to install and maintain. We expect that server-based services will

increasingly become easier to configure on individual desktops--as computing power increases and market forces drive software design towards more simplified and automated installation. Indeed, personal WWW servers are already becoming quite common on consumer operating systems such as Windows 95 and the MacOS, and many X server software packages for the PC contain applications that enable one to set up their machine as a WWW server, telnet server, ftp server, etc.

IMPLEMENTATION

Estimates of access to network bandwidth and computing power typically available to ecological scientists constrained our potential solutions, as did the current immaturity of available groupware solutions. Some of the more sophisticated groupware packages required major investments of time and money to use effectively, as well as installation of specialized software on each potential client machine—thereby disqualifying them as solutions for NCEAS' needs. Furthermore, some groupware packages require that all client systems exist within a LAN, and are not yet effective for integration over a wide-area network (WAN), especially large, public-access networks like today's Internet.

Design goals

Instead of using proprietary groupware approaches, we chose to adhere to open standards, and turned to the Web as our mechanism for delivering groupware solutions. The use of standard Internet services provided us with maximal interoperability--enabling anyone with Web access to use our services, regardless of whether they are on UNIX, Macintosh or PC systems. We identified several achievable features that we believed would provide substantial benefits to distributed workgroups of ecologists. These included: assurance of privacy for the groups' materials; simplification of storage and access to the materials; exchange of richly formatted items, such as graphics and proprietary data formats; and facilitation of rapid exchange for upload and download. We accomplished all our services using freely available server software, and require only email clients and a graphical Web browser (ideally one supporting frames) for client access.

Privacy

We provided privacy to each working group via access control mechanisms that are available on many Web servers. The specific mechanisms vary from one server package to another, but these all essentially involve some systems-level configuration of an authorization file allowing computer or user-level access to specific areas on the Web server (see section 5 of the WWW Security FAQ-- <http://www.w3.org/Security/faq>, or NCSA's user authentication tutorial-- <http://hoohoo.ncsa.uiuc.edu/docs/tutorials/user.html>). We use single accounts for each working group, with everyone in the group sharing a password. This is not ironclad security (e.g., a firewall), but it provides a sufficient measure of confidentiality for scientific working groups. The user authentication mechanism also leads to a sense of closeness among the collaborators due to the focused content of the prescribed area to which all the participants are contributing.

Single, authoritative archive

We identified an archival function as critical for the groups' virtual work areas. One of the difficulties for any scientist is filing email and their attachments, shared working papers, intermediate analyses and data sets, etc., and tracking these with respect to updates and revisions.

We provide a single repository for all these items, so that group members can login to their private Web area whenever they want to review information or download the latest update to the data. The onerous task of individually filing group work is eliminated. Instead, one individual from each group usually becomes the 'steward' for information placed within the private areas, and coordinates the group with respect to versioning of preprints, intermediate data sets and analyses, etc.

Closed email list

An important component of the virtual working groups is the creation of a private email list. Each collaborative group at NCEAS has its own email list. These lists only accept postings from list members, protecting the group from unwanted mailings. The list is centrally managed, so that modifications are immediately in place, and individuals don't need to maintain their own separate email aliases. Also, rather than expecting each scientist to individually file these messages, we maintain a complete archive of the mailings within each group's private work area. These archives are presented in a threaded format on the Web, so that mailings related to the same subject can be grouped together. In this way one can follow a discussion rather than browsing through the entire archive holdings for information about a topic.

We chose the procmail package with SmartList extensions (<ftp://sunsite.informatik.rwth-aachen.de/pub/packages/procmail>) for email list creation and management on our systems. It is freely available software that will run on any UNIX platform, but it can require some significant configuration of your email server. We use the procmail/SmartList combination because it is more flexible and scalable than another popular and free mailing list management package called 'Majordomo' (<http://www.greatcircle.com>). We used the software package hypermail in order to provide threaded, HTML-formatted messages within each groups' private Web area. This free software runs on UNIX machines (<http://www.eit.com/packages/hypermail/hypermail.html>), and translates UNIX mail-formatted messages into HTML documents cross-referenced by subject, author, and date for threading. The threading is a convenient feature, commonly found on network news readers, that is also quite handy for grouping email, particularly by subject or sender. Ideally, correspondents pay attention to the subject line of their email, so that messages will file properly within a subject thread, or spin off on a new topic.

Rapid exchange of graphics, data, and text

Email software has grown in functionality over the past several years to include easy attachment and automatic decoding of any type of file, including non-textual materials such as graphics and executable software. Still, it is not yet efficient to send data files via email, since correspondents frequently find that they cannot decode one another's attachments. Also, email is typically delivered to an area of the mail server that may not be equipped to handle potentially large files, such as raw data or detailed image files. Both of these issues should soon resolve towards easier exchange of files via email, as email software complies more closely with the Internet MIME standard specifying how attachments are to be embedded in messages, and servers are configured to deal with large transient mail spools. But as of 1998, email is not entirely satisfactory for non-text file exchange.

Web browsers are also deficient for file exchange, due to their current lack of flexibility in uploading files. Capability to upload is built-in to the http specification that Web browsers use (PUT or POST commands), but current browser software tends to limit implementation of these functions to Web forms, and not allow unrestricted upload of an arbitrary file. For this reason, we

resort to the old Internet standard *ftp*, or file transfer protocol, for file upload. NCEAS creates special areas that allow working groups to accomplish ftp uploads of data to our system. We then move these files over to the private areas following the instructions of the individual groups' data stewards. We find that Web browsers serve quite well for downloading files. NCEAS' staff provides some minor html-embellishment to make presentation of the files more informative and attractive for prospective downloading by group members.

Use of ftp also enables us to avoid the difficulties with email exchange of attached materials. For example, users of Eudora[®] software frequently employ (often unwittingly) an encoding format called 'binhex' to their attachments. Unfortunately, 'binhex' is virtually unsupported outside the Macintosh and Eudora realms, so non-Eudora users are often unable to decode these types of attachments. By using ftp, our scientists can upload any type of file, and then use their Web browser to download any type of file. No special encoding--binhex, base-64 or other--is necessary, so no decoding difficulties arise.

We also sought a way in which scientists could exchange fully-formatted documents including text and graphics, without having to worry about everyone having the same package or version of software in order to read the file. We chose the portable document format, PDF[®] from Adobe[®] (<http://www.adobe.com/prodindex/acrobat/adobepdf.html>) to provide us with a cost-free, platform-independent way to accomplish this. Adobe's Acrobat[®] Reader is freely available for all the major UNIX, PC, and Macintosh operating systems. NCEAS' staff currently facilitate the translation of Word[®], WordPerfect[®], etc. documents into PDF by using Adobe's Acrobat software—which is not cost-free as opposed to the Acrobat Reader. The derived PDF files go directly onto the Web, where they are available for viewing by anyone with the Acrobat Reader software. One drawback to PDF is that the files are only suitable for viewing or printing—editing is not possible without additional costly software add-ins to Acrobat.

FUTURE FUNCTIONALITY

Several previously described standard Internet services were used to create 'virtual working groups' on behalf of the scientists collaborating through NCEAS. These methods had the use of the Web at their core--as a centralized and authoritative repository for email transactions, and a restricted-access location for the exchange of rich text, graphics and raw data. There are a number of other groupware functions, however, which we were unable to effectively implement given the constraints of today's networked computing environment. I will discuss these briefly here, as a preview to what ecologists can expect to become commonplace services within the next several years.

Real-time application sharing

All the capabilities described above for NCEAS' virtual working groups involve asynchronous interactions: individuals upload or download data, send messages, etc., and other participants access this information through the Web when it is convenient. But perhaps the most requested capability that we cannot yet support is that of real-time application sharing. This would enable multiple individuals to jointly view and control a program, with envisioned usage primarily that of collaboratively working through analyses using packages such as SAS[®] or MATLAB[®].

Luckily, most robust scientific and analytical packages run optimally on computers with UNIX operating systems, where the X Window System, or X11, is the standard for graphical display (<http://www.opengroup.org/tech/desktop/x>). X Window is built on a platform-independent, client/server model, so one is able to run X-based graphical applications (clients) on remote computers as long as one's local computer has network connectivity and X-server software. X and

the Motif interface comprise the standard GUI for most UNIX systems, and are available as add-on software for PC's and Macintoshes. X-based applications running on NCEAS' big UNIX servers enable remote scientists to login to our systems and run applications as if they were on-site.

The primary complication to running X applications over today's networks is bandwidth: 'best effort' service of current TCP/IP networks frequently leads to unsatisfactory performance when running over a long-haul network, such as the Internet. As described in Table 1, however, network bandwidth to scientists' desktop computers is likely to increase substantially over the next several years. This will create a situation in which remote logins to powerful computers running X Window applications will provide extremely good performance and potential for interaction. Nevertheless, X Window has already proven quite convenient to our remote collaborators in running less graphically intense jobs on NCEAS systems. Since UNIX is a robust multiprocessing and multi-user operating system, NCEAS' servers can support numbers of remote individuals simultaneously accessing and running jobs, each with graphical interfaces to applications via X Window.

There are several groupware implementations of X, which enable multiple, remote users to share a graphical session. Notable among these is the freeware package, xmx (<http://www.cs.brown.edu/software/xmx>), which runs on UNIX systems. There are also some professionally supported X multiplexors which we did not test. Xmx allows multiple accounts to access the same X Window session, and to pass control of the mouse and keyboard among the participants sharing the application. Unfortunately, the program does not currently support connections to PC's running X server software. PC support is a planned feature, and may be available by the time this article is in print.

Multimedia Teleconferencing over the Internet

Multimedia teleconferencing comprises several valuable services that will greatly facilitate scientific collaboration over the Internet. Of these functions, NCEAS' scientists could benefit greatly from a shared whiteboard: 'live' document-sharing, in which multiple participants can draw, modify, and annotate shared graphics and text in real time. Ecologists would ideally like to accomplish this with audio contact over the network, to easily converse about the shared display (and without incurring long-distance telephone charges!). It would also be nice if this were a multi-point service, rather than 'point-to-point' (only two participants). The concept of document sharing can be extended to that of application sharing--allowing for collaborative work in popular desktop applications which often are not compliant with the X Window protocol described in the preceding section. Ultimately, we expect these features will be integrated with full multi-point live video contact and bi-directional file transfer.

These are all services for which recently defined standards promise a number of interoperable products in the near future. The relevant specifications are the International Telecommunication Union's (ITU) T.120 standard for real-time data conferencing, and the H.320 series for video conferencing (for updates on the status of these still-developing standards, see: <http://www.imtc.org/imtc>). The T.120 standard provides a common specification for application sharing, and future applications built in compliance with the standard will be a potential alternative to the shared X solutions described above. T.120 also describes how multi-point data conferences can be achieved in a vendor-neutral way. Similarly, H.320 provides standards for how audio and video are to be compressed and delivered over varying bandwidths in a vendor-neutral way. At the present time, however, there is lack of interoperability among most vendors' offerings, so that, e.g., PC users running a Window's conferencing application might be unable to communicate with Macintosh or UNIX users.

We have tested several conferencing solutions at NCEAS, but these all have some drawbacks--relative to application instability (essentially, *beta* software that crashes or performs erratically), unsatisfactory performance over a long-distance network connection due to insufficient and/or unreliable bandwidth, and lack of interoperability prohibiting a broad user base running on PC's, Mac's and UNIX boxes. Among recent releases, Netscape's Conference comes closest to providing a cross-platform solution, with support for point-to-point shared whiteboarding and audio communication for PC/Mac/UNIX. Microsoft's NetMeeting currently only runs on PC's with Windows 95 or NT.

The emergence of the ITU standards coupled with growing network bandwidth to the desktop will catalyze unprecedented growth in network-based multimedia teleconferencing over the next several years.

CONCLUSIONS

At NCEAS, we currently use the WWW to enable groups of scientists to collaborate on ecological research projects. By providing a single, private area for each group to post and discuss data and results, we relieve scientists of the task of individually tracking and filing these items. The private email list enables individuals to quickly communicate with a potentially changing group membership, again with the convenience of knowing that all messages will be archived in threaded format and available through the Web at any time for easy reference. Use of the PDF format enables scientists to share works-in-progress with full formatting and graphics, regardless of what specific software programs were originally used to generate the materials.

NCEAS' powerful UNIX servers support multiple users running multiple applications--capabilities not yet available on common desktop computer operating systems. By using X Window, remotely located scientists are able to accomplish analyses on our systems while working with a full graphical interface. Current unpredictability of network bandwidth, however, prevents remote users from running applications over the network on our systems as effectively as they might from a workstation on their own desktop. This annoyance will disappear when the next generation Internet provides ample bandwidth to academic and research communities.

Internet technologies themselves are still rapidly evolving. As network bandwidth increases and desktop computers continue to grow more powerful, more complex and effective services become possible. Standards groups continue to define and extend how interoperable solutions can be promulgated over the Internet. Groupware is one area that is likely to benefit greatly from these developments, since these solutions can be very demanding in terms of bandwidth and computational power, and become far more effective if they are capable of running on multiple platforms.

Two technology trends stand out in particular with regards to possibilities for enhancing scientific collaboration. First, the client-server model for delivery of most Internet services will continue, but individual desktops will become more capable of delivering server-side solutions as the relevant software becomes simpler to install and maintain on hardware that can deliver satisfactory performance while handling multiple clients. This will be especially true for smaller working groups seeking modest levels of service to facilitate close interaction. Second, the trend towards full multimedia teleconferencing will continue over the next several years, as standards solidify and guaranteed adequate bandwidth reaches the individual desktop.

ACKNOWLEDGEMENTS

I would like to thank Matt Jones for hours of stimulating discussion on how to provide networked computing services to ecological scientists. I'd also like to thank the conveners of this

workshop--Bill Michener, Jim Gosz, Art McKee, and John Porter--for giving me the opportunity to present some of the solutions we've adopted at NCEAS. Finally, I want to express my appreciation to the scientists coming through NCEAS who were willing to discuss and test new approaches to doing collaborative work in the ecological and environmental sciences.

This work was funded by the National Center for Ecological Analysis and Synthesis, a Center funded by the national Science Foundation (Grant #DEB-94-21535), the University of California – Santa Barbara, and the State of California.