AN INTEGRATIVE INFORMATION FRAMEWORK FOR ENVIRONMENTAL MANAGEMENT AND RESEARCH

D. D. Cowan, Dept. of Computer Science
dcowan@csg.uwaterloo.ca

T. R. Grove, Computer Systems Group
trg@csg.uwaterloo.ca

C. I. Mayfield, Biology Department
mayfield@sciborg.uwaterloo.ca

R. T. Newkirk, Urban and Regional Planning
newkirk@watserv1.uwaterloo.ca

University of Waterloo;Waterloo, Ontario; CANADA N2L 3G1

D.A.Swayne, Computing and Information Science,University of Guelph
Guelph, Ontario; CANADA N1G 2W1
dswayne@sciborg.uwaterloo.ca

ABSTRACT

Effective environmental management and research require environmental and socio-economic data, information and knowledge to anticipate and predict the impact of development. However, current information, while often extensive, is fragmented, inconsistent, under-utilized, and often inaccessible. These factors have lead to "information gridlock" where the data and information is available but requires inordinate amounts of time and expertise to begin the process of acquisition and assimilation. There is a desperate need for an integrated environmental information system containing both a knowledge base and decision-support tools to help manage the acquisition process. The University of Waterloo in partnership with several organizations from both the public and private sector has embarked on a large-scale project to develop integrated information systems for the management of human use of the environment. This paper will describe some aspects of the system architecture and the approach to implementing this architecture.

INTRODUCTION

Managing mankind's use of the environment is one of the most important social and political issues of the next several decades. The current environmental stresses on planet Earth are "unsustainable" and fundamental changes to the way in which we manage the use of our entire habitat are essential.

Environmental issues cover many concerns including global warming, ozone depletion, contamination of local groundwater and preservation of fish habitats. The American Scientific Community has chosen global climate change [Gershon 1993] as one issue of concern, but this leaves many other extremely important environmental issues, such as the management of urban regions, as key questions to be addressed. With half the world's population urbanized by the year 2000, sustainable management of urban regions will be a key environmental issue of the next century.

One of the major obstacles to improving the way we handle all manner of environmental problems is the lack of easily accessible and useful environmental and socio-economic data and information. There is no lack of this type of data and information, but it is

inconsistent, stored in diverse locations on both paper and in various incompatible machine-readable formats, and on many different computer platforms. Even different agencies of the same levels of government often use incompatible data storage technologies and formats. Accessibility is a particularly serious problem in urban regions, because environmental and socio-economic data and information are often held by different and often overlapping levels of government, as well as by various categories of consultants and public interest groups.

This lack of accessibility has led to "information gridlock" where very large quantities of environmental and socio-economic data and information are available, but cannot be effectively accessed or processed to yield insight into the management of mankind's impact on the environment. Overcoming "information gridlock" should be a significant step toward reaching the goal of sustainable development.

The result of this "information gridlock" is that much time and effort is employed at significant expense, redeveloping and recreating data, rather than efficiently applying existing data and information to environmental problems. The need to resolve this gridlock problem has been recognized by the United Nations, the Brundtland Commission and many national governments.

A BRIEF OVERVIEW OF CURRENT INFORMATION SERVICES

Computer-based tools already exist to assist with organizing and accessing of data. Users of the Internet[Krol 1993] are familiar with such applications as "FTP" and "Archie"[Emtage 1993], Gopher[Alberti 1992] and Veronica[Foster 1993], "WAIS" ("Wide Area Information Servers")[Kahle 1989] and "WWW" ("World-Wide Web")[Berners-Lee 1992]. Most of these tools require (at least) a moderate degree of computing expertise in order to be used effectively. For neophytes to the Internet, the learning curve for these tools can be overwhelming.

Others have identified these limitations: several projects are underway to produce computer-based information systems that are focussed on the problem of environmental information systems (or "EIS"), including "CIESIN" ("Consortium for International Earth Science Information Network")[CIESIN 1993], "ERIN" ("Environmental Resource Information Network")[Slater 1992], "GENIE" ("Global Environmental Network for Information Exchange")[Newman 1992], and "Lamont"[Menke 1991].

The goals and interim results of these efforts seem to concentrate on accessibility of information (location, client software), and less on the information management problem. For an EIS to be successful, it must also provide information management tools: in many cases, this entails acting as an information filter, as well as an information repository.

OBJECTIVES OF THE EIS FRAMEWORK

A fundamental goal of an EIS is to enable its users to locate, acquire and process information relevant to a problem, and then present results in a meaningful fashion. These users range from "domain experts" (planners, environmental engineers) to the general public, and they often lack the computing expertise needed to perform the necessary tasks. An important

component of an EIS is to provide geographic-based mechanisms for finding and obtaining information, and facilitate using that information to solve problems.

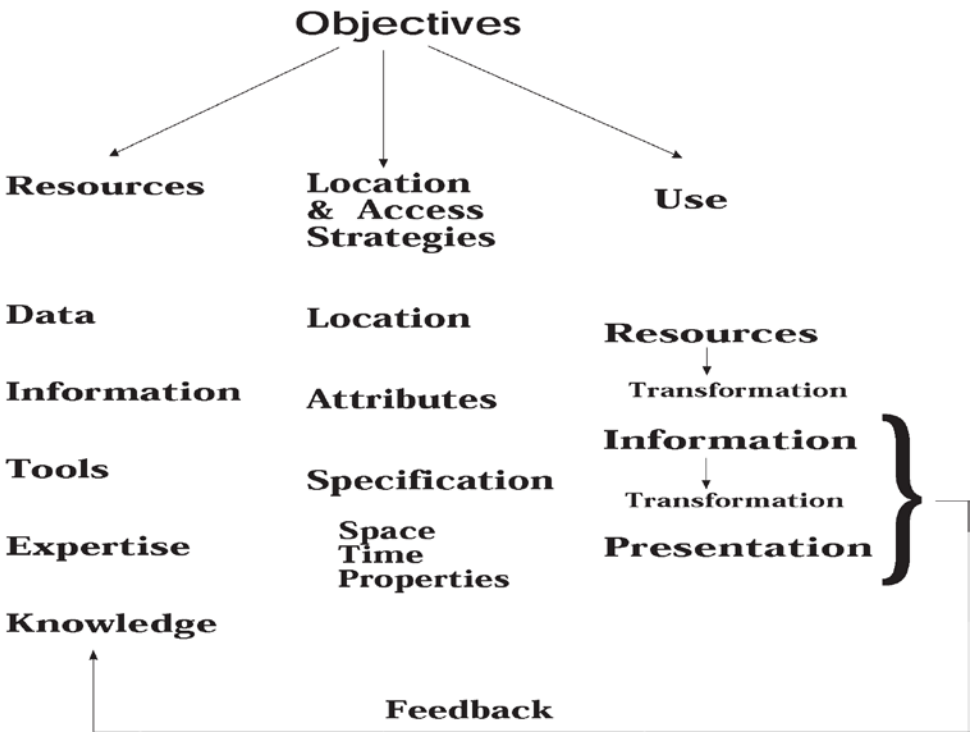

Figure 1: An information architecture

To some degree, the functions performed by an EIS model that of a consultant: the gathering, processing and presentation of information. In many cases, the information gathered by a consultant already exists, and simply needs to be accessed. Informally, an EIS strives to provide users with the ability to be their own consultants.

Figure 1 shows the general framework of an EIS architecture. The **Objectives** represent the problem to be solved or the decision to be made. **Resources** are the input to the solution or decision process, and **Use** represents the actions or **Transformations** that are to be performed on the resources. The **Location and Access Strategies** column represents the component of the EIS that enhances the accessibility of the resources and facilitates improved integration of information and processes.

The "feedback loop", where derived information and knowledge is returned to the resource base for subsequent use, indicates a mechanism whereby the information framework provided by the EIS can be used to magnify and amplify expertise. As information and knowledge are created, they can be made available as a resource to other users.

The designer of an EIS must consider two diverse classes of user: experts who are scientifically sophisticated and have a long-term focus on narrowly-defined problems; and integrators, such as municipal planners, who have momentary focus on a problem, but have many such problems.

Locating and Accessing Resources

It is estimated that 35% to 40% of consulting costs are allocated to acquiring resources such as data, information, expertise and knowledge (where "knowledge" is viewed as "encapsulated expertise" comprised of encoded data and rules). Expertise arises in several categories, including: technical (focussed knowledge in limited domains); integrative (broad base of general knowledge and experience in its application); and adaptive (cross-disciplinary knowledge and experience). The resources are accessed independently and repeatedly by consultants and practitioners.

Additional costs in information-gathering are incurred because of the computing expertise that is required to work within the existing ad hoc framework. A primary technical goal of the new EIS framework is to reduce this cost by providing a flexible infrastructure to expedite acquisition and utilization of information. The qualitative improvement over "raw" data repositories will achieved through the use of partially-automated location and access strategies.

Since the fundamental purpose of the EIS is to improve the ability of the user to acquire and use information, the access strategies that are embedded into the system should eliminate the need for substantial computing expertise and should hide the details of data access.

The user view

The user view presented here is solution-oriented: the use of the system is motivated by the need to solve a problem. In this context the goal of the EIS is to improve the "connection" between **Resources** and **Use**.

In essence, there are only a few types of questions that are commonly asked in an environmental context. For discussion purposes, the focus will be on the question "What is the effect of doing *X* on a volume *Y*,*t*?"; which is representative of commonly-asked questions. *X* denotes a physical activity or effect, and *Y* denotes a geographic region with a possible temporal component. Repeated (iterative) application of this class of question is commonly used to determine if regulatory or performance criteria can be met.

An EIS can help to answer this kind of question by identifying tools to help to define and resolve *X*, and by helping to specify or select other resources related to *Y*. For example, *X* might be "increased rainfall", and a specification of *Y* might be textual (consisting or words or phrases that describe the region) or a polygonal description of the area. If a specific type of study is to be done, the EIS can help to locate appropriate information. Conversely, there may exist a dataset that must be analyzed: in this case, the EIS can assist in finding an appropriate tool. It is often the case that there is a "feedback loop" between these two situations: availability of data may motivate the selection of a particular tool, which requires additional data, and so on. It is a practical reality, because of expense, that many environmental assessments are done with the available data, rather that completely suitable data.

DESIGN OF ACCESS STRATEGIES

Given that there already exist numerous data repositories and collections of tools and modelling systems, the major contribution of the proposed EIS framework is to facilitate access to these items. It is apparent that the existing telecommunications infrastructure initially is sufficient to build some prototypes of an EIS: hence the access strategies concentrate on the

organization of the data and tools and on standardized access methods. Paramount in the organization is the ability to encode queries and data geographically.

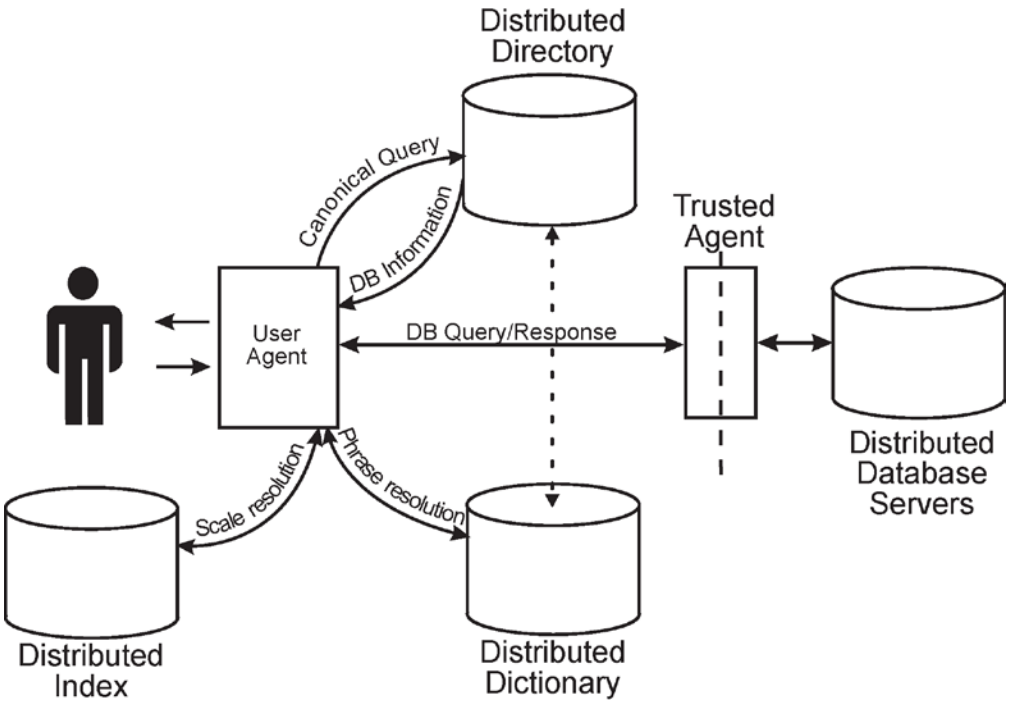

Figure 2: Framework for an EIS

Figure 2 is a representation of the proposed EIS framework. There are a few components in the EIS framework that require special attention: a dictionary service that contains information about geographic references; a directory of available databases, tools and knowledge bases; a directory index used to resolve database scale and resolution problems; and a trusted-agent mechanism to control access to databases.

These services will be available for use via servers located on a high-speed wide-area network (such as NREN[Aiken 1992], CANARIE[CANARIE 1993] or the Internet). Access mechanisms for a particular item listed in the directory will be defined in its directory entry: although it generally will be accessible via the same network as the framework services, this need not be the case.

The descriptions of the services here are organized by function. In reality, it is likely that the implementation of the framework will be done on an organizational basis. Hence, the design of the framework must encompass fully distributed implementations. The X.500 standard[Bumbulis 1993] for directory services may provide an important standardized base. An important aspect of this standard is its facilities for user authentication, providing flexible access control and security for the databases.

Clearly, the potential scope of the framework described here is global. It is critical that the first stages in the development of the framework concentrate on laying the foundation of the framework and creating a suitable "recipe for change". The example of the evolution of the Internet may be instructive in the task. The original sponsoring agency of the Internet (as it became to be known) specified only key protocols and implemented only small prototypes.

Many applications were developed independently, and much of the growth was achieved incrementally by improving the availability of developing standards and applications.

Dictionary services

A dictionary service contains information about geographic references. For example, in the query "what is the effect of toxic spills in Bloom County?", the reference "Bloom County" would yield a suitable polygon description of the area. "Suitable" is a complex decision in this case. The obvious, exact polygon for a particular reference might not be appropriate in some circumstances (for example, computational complexity caused by fractal-dimension effects on boundaries that follow geographic features). Hence the user should be provided with the opportunity to review the "dictionary definition" of an item before its use.

The ideal dictionary service is a knowledge-based system that can provide advanced searching and pattern-matching capabilities. However, simpler mechanisms will fit easily into the framework, and can be replaced in a modular fashion as new and improved components are developed.

An important aspect of the framework is the ability to capture new knowledge as it is created. This is especially important for the dictionary services. For example, if a new textual geographic reference is defined (by a user), it will be of value to all users if that new reference can be made available. This update facility will require the use of trusted agents or some similar mechanism.

Directory services

The directory service provides information about databases, knowledge-bases and tools. A query might contain a request for a list of databases that contain information about a geographic region, or a request for a list of tools that pertain to a given keyword. Queries are pattern-matched against the entries in the directory, resulting in a list of matching entries. For example, in the previous sample query, the keyword phrase "toxic spill" might be used as the key to search for tools, and the polygon description of "Bloom County" (as provided by the dictionary search) would be the key for the search for databases.

It is imperative that a standard representation (a *canonical form*) for each field in the entry is available. The directory entry indicates whether the canonical form of a field is available: if it is not, the owner of the database is expected to provide transformation functions that can be used to convert the private representation to the canonical form. These functions could include textual descriptions, formulae or computer programs. Whatever the case, the transformation functions would be made available via the framework. It then becomes the responsibility of the user agent to ensure that a query contains the appropriate data representation.

Directory entries must be sufficiently flexible to reflect the variations in the items that are to be represented. Each entry would have the following types of fields:

- temporal. Some items are applicable to restricted time ranges, and such information should be recorded in the entry.
- spatial. The canonical form (in the directory) for spatial information should be polygons. For a database or knowledge base, this is the area covered by the database; for a tool, it represents any restrictions on the regions for which the tool can be used.
- data specifications and attributes. This information describes the characteristics of the data that is contained in the database: for example, soil data or groundwater information. The

keywords used in the database descriptions are from the same domain as the keywords that describe the tools, thus allowing searches to be restricted to databases that contain information suitable for use by specific modelling tools.

Clearly, items listed in the directory may contain other classes of information of interest to users. The design of the directory services must accommodate searching of such indeterminate information. In general, a query should be able to specify matching criteria for any of the fields in the directory entry.

As noted, the directory services probably can be based on the X.500 directory services model. Directory users will forward search requests to a directory user agent that will perform the search and produce a list of items that are (or may be, depending upon search criteria) of interest. Subsequent access to databases can be direct or via a "trusted agent" that accepts and verifies requests. Such a technique provides full access control to database owners, if they so choose. Thus, the responsibility and control of databases is outside the domain of the EIS framework.

Trusted agents

Database ownership, security and access control are of prime importance to data providers: recognition of the proprietary nature of the intellectual property contained in databases must occur if the commercialization of the framework is to be considered. Hence, it is important, at the outset, to incorporate standards for access control, access charges ("pay for service") and similar issues. The trusted-agent mechanism is designed to accommodate such requirements. The trusted agent can act as a "network gateway" for the database, isolating the server from the public data networks, tracking usage and calculating access fees.

The security mechanisms embodied in the trusted agent protocols can range from simple (a null process, if the provider wishes to allow unrestricted access) to complex (a *Kerberos*-style [Steiner 1988] authentication system), or some measure between these two. In general, the trusted agent is just a process that intervenes in all accesses to a database, and provides the "hook" for all security and access control mechanisms.

It is envisioned that the trusted agent component of the framework will consist of the protocol for the messages that users transmit to the trusted agents to request access to the data. The specification of the protocol is the responsibility of the framework, but the actual implementation of the agent is the responsibility of the data provider. This division of function gives the data providers the assurance that it is their code (and only their code) that is actually "touching" their database. To simplify the process of implementing the framework, sample implementations of a trusted agent should be made available to data providers.

Processing user queries

The question "What is the effect of doing *X* on a volume *Y,t*?" is the user's view of a question. This question can be decomposed into simpler query primitives that are easier to implement with typical information-retrieval and database techniques:

1. List all information about *X* that is contained in *Y,t*.
   Region Y may be a jurisdictional name or a geographic polygon (for example, specified with an interactive map-based query tool). If it is a name, it must be transformed into a polygon (i.e., the canonical form for spatial information). The dictionary service will provide such transformations. Once a region-reference has been resolved, the directory is

searched, using both the area polygon and *X* as keys for the search. *X* is a keyword that describes a geo-political attribute of a region.

2. List all of the tools or knowledge bases that pertain to *X*.
   The directory contains entries describing known modelling tools and knowledge-bases. Associated with each entry is a list of keywords (of the same domain as *X*) that are used to define the applicability of the tool. The result of a query is an enumeration of tools and knowledge-bases that may relate to *X* in some fashion.

All user queries can be represented with a logical combination of these or similar primitives, and hence, the EIS framework must support processing of such of primitives.

The directory index

The search process implied by the first class of question in the previous section yields both intersecting and enclosed polygons. The enclosure case can cause matches with too large an area, leading to false (or meaningless) matches and an information "overload". Consider, for example, a directory containing entries for databases about "Bloom County" and "the world". The polygon for "the world" will match a search for "Bloom County", but this result is of little value.

Thus, to be effective, the search must be refined in some fashion. Since it is likely that the database described by the world-wide polygon consists of smaller units, a scaling operation can be applied to the search. The database index provides information about the composition of polygons in the database directory. Using the index, the constituents of an enclosing polygon can be examined to determine if they form a better match for the enclosed (search target) polygon. Constructing an index for an existing database represents an important value-added service.

OBSERVATIONS AND CONCLUSIONS

It is a simple task to describe the ideal EIS: implementation of such a system is not so easy. Practical realities require that consideration be given to the real-world problems of incorporating existing geographic information into any computer-based EIS (the data legacy problem).

The proposed EIS framework does not specify an end product: it specifies a base that can be achieved with existing technology, and provides a standardized path for incremental enhancements to the system. It is important to recognize that there is no ideal system: any successful EIS will, by its very nature, be a dynamic entity.

The most difficult step in the implementation of the EIS is the "first step". In order to begin to provide a useful system, existing data owners must be motivated to provide access to their data, and applications must be created to make use of data that is unlikely to be in any standardized form. This is clearly a "chicken and egg" problem: framework-conformant applications cannot be developed until framework-conformant data is available, and such data will not be made available until data owners can get something in return for their participation in the EIS. Additionally, potential users of any system will not be interested until there is some "critical mass" of available information that will justify the new system's learning curve.

The problem of "jurisdictional overlap" of data, where several independently-maintained databases exist for a given geographic region, is already a problem for

practitioners. For a computer-based system, this overlap will be compounded as more and more information is made available through the framework. It may be necessary to incorporate some notion of authority into the EIS, in the sense of identifying the primary source versus secondary sources of some particular set of data.

Implementation of the EIS framework is divided into two categories: technical (the computing technology and environmental knowledge); and organizational, which is the problem of motivating database owners and application providers to adhere to the framework.

Solutions to many of the technical problems can be obtained with existing technology. However, the EIS can, and will, be improved in the future with better technology (faster networks, sophisticated knowledge-based front-end applications to assist searches, better environmental applications, and so on), but the fundamentals of the framework can be designed and implemented now. Note that it will be feasible to geo-code only the directories, and not the databases themselves (although database owners should be encouraged to do so).

Solving the organizational problem is not as straightforward. Attracting users to any new system requires that it provide qualitative and quantitative improvements to the status quo. There are two immediate benefits to EIS users: information about databases and access requirements will be centralized; and the productivity of existing experts will be enhanced. As the EIS expands and becomes more sophisticated, additional benefits, such as accessibility to "less-than-expert" users and elimination of the need for computing expertise, will evolve.

Motivating information providers is more complex. A small, select group of agencies should be selected as "partners" in the construction of a prototype of the EIS, using a small test-case to prove the concepts. Over time, the framework can be "scaled up" and involve additional application developers and data proprietors. This user-directed, "ground up" approach is in contrast to the top-down approach that has been used elsewhere. To encourage database providers (and indeed, users), the cost of joining the framework must be minimized, and the value of joining must be maximized. Retaining control of data is another key component in motivating data owners.

The approach proposed in this paper would probably languish in a research laboratory unless key users are motivated to adopt this structure both internally and as an access method for external users. Finding appropriate partners to adopt the technology is as important as the creation of the original concept.

REFERENCES

Aiken, R. et al. (1992) NSF Implementation Plan for Interagency Interim NREN. Available via Gopher from gopher.es.net in NREN/NREN documents/impl.txt.

Alberti, B. et al. (1992) The Internet Gopher Protocol. Technical report, University of Minnesota, Microcomputer and Workstation Networks Center. The Gopher FAQ is available via anonymous FTP from rtfm.mit.edu in /pub/usenet/news.answers/gopher-faq.

Bumbulis, P.J., Cowan, D.D., Durance C.M. and Stepien, T.M. (1993) An Introduction to the OSI Directory Services. *Computer Networks and ISDN Systems*, to appear.

Berners-Lee, T.J. et al. (1992) World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*, 2(1):52--58, Spring 1992. Published by Meckler Publishing, Westport, CT, USA.

CANARIE (1993) *CANARIE: The Canadian Network for the Advancement of Research, Industry and Education*. CANARIE Inc.; 401 Laurier Avenue West, Suite 1120; Ottawa, Ontario, Canada K1P 6H5. The mission statement for the CANARIE project.

CIESIN (1993) CIESIN Mission and Activities Overview. Briefing prepared for the United Nations Sustainable Development Network, February 1993.

Emtage, A. et al. (1993) *Archie MAN page*. Available via anonymous FTP from ftp.cc.mcgill.ca in /pub/Network/archie/man/archie.man.txt.

Foster, S. and Barrie, F. (1993) *Frequently Asked Questions about Veronica*. Available via Gopher from veronica.scs.unr.edu in veronica/FAQ.

Gershon, N. and Miller, C.G. (1993) Dealing with the data deluge. *IEEE Spectrum*, 30(7), July 1993. This issue contains the special report "Monitoring Global Climate".

Kahle, B. (1989) Wide Area Information Server Concepts. Technical Report TMC-202, Thinking Machines Corporation, 1989. Available via anonymous FTP from quake.think.com in /pub/wais/wais-concepts.txt.

Krol, E. (1993) *The whole Internet: user's guide and catalog*. O'Reilly.

Menke, W. et al. (1991) Sharing Data over Internet with the Lamont View-Server System. *EOS*, 72(38):409--416, September 17 1991.

Newman, I. et al. (1992) GENIE -- The UK Global Environmental Network for Information Exchange. In *Computing in the Social Sciences 92*.

Steiner, J.G. et al. (1988) Kerberos: An Authentication Service for Open Network Systems. In *Winter USENIX*, Dallas, Texas, U.S.A..

Slater, W. (1992) ERIN Concepts. Available via Gopher from kaos.erin.gov.au. Contact the author at wayne@erin.gov.au.