AARNet

ARCHIVES WORKING GROUP

FINAL PROJECT PROPOSAL

July 1991

1. Introduction

This report is the work of the AARNet Archives Working Group. It proposes the establishment of a single, large computing system to service the information retrieval requirements of the AARNet community.

The goals of the project are to reduce the volume of duplicate file transfer requests to international locations, and to improve the on-line information services currently available within AARNet.

2. History

The submissions for 1991 AARNet Project funding included a proposal from Stephen Cliffe of the University of Wollongong, that addressed many of the issues related to the location, retrieval and archiving of information, data and programs within AARNet.

In considering this proposal, the AARNet Advisory Board Project Evaluation Sub-Committee recognised that these issues were of such importance that the working group formed to prepare the original submission be funded to develop a more detailed project description.

The members of the resultant AARNet Archive Working Group are:

Stephen Cliffe, University of Wollongong
Craig Warren, Deakin University
Michael Henry, University of Tasmania
Mark Prior, University of Adelaide
Ross Cartlidge, University of Sydney
Peter Elford, AARNet

A meeting was held on 12th June, 1991 at the ANU, Canberra to prepare the following project proposal.

3. Background

The introduction of AARNet into the Australian Academic and Research sector has made it possible for members of the university and CSIRO community to gain ready access to a wealth of data, programs and documents stored in electronic form on computer systems throughout the world. Although the ability for any user with a connection to AARNet to be make use of these resources is very attractive, it presents two major issues to the users and engineers of AARNet.

The first is a network engineering concern. The bulk of the information available over the global networks is not located in Australia. This means that any information brought into Australia must transit the international satellite link between Melbourne and the United States. Unfortunately, in the absence of any other mechanisms, if more than one user wishes to obtain a copy of an item, multiple copies of this item will be transferred into the country.

If file transfer was a small contributor to the total volume of traffic on this link then this would not be a major engineering issue. However, as the graphs of international traffic show (attached), file transfer not only accounts for the largest percentage of traffic volume, it is growing at a much faster rate than any of the other network applications. Such explosive growth is extremely difficult to engineer solutions for, partly because of the difficulties in estimating the extent and duration of the growth, and partly planning for adequate resource provision to support the potential demand.

The second problem is essentially one of user education. There are several specific issues that need to be addressed within this broad topic;

a) Information Resource Identification. Frequently users are unaware of the scope of the information resources available to them;

b) Location of Information Resources. Users may be aware that some information they require, be it data, programs or documents, is on-line but do not know the name and/or network address of the computer system(s) upon which it is located;

In many cases, copies of foreign information resources are maintained in Australia, but users are unaware of them and continue to transfer data from sites in the USA, Europe and the rest of the world. This problem is further complicated by the fact that local (Australian) copies of information resources are sometimes out of date, and similarly difficult to locate;

c) Poor Directory Tools. The existing tools for addressing the identification and location problems (a and b above) are crude, unwieldy and inefficient. Users are therefore not prepared to spend the time and effort to find the "closest" or "most network efficient" location of information;

d) Announcements. New versions of software, technical reports, new data sets, etc. are frequently announced to the networked user community via electronic mailing lists and network news bulletins. These postings generally only cite the original source location of the item, and therefore generate a rush of identical requests to the same, usually international location, from the users that read them. Ideally, one copy should be brought to Australia and then made available to the rest of AARNet community nationally;

e) Poor Access Tools. The standard method of information transfer is based on FTP (File Transfer Protocol) and the "anonymous" user concept. FTP itself, and the "anonymous" guest account mechanism, are not ideal end user interfaces for the distribution of information. They are however, widely supported and available on all hardware and software platforms;

f) Poor Performance. Because some sources of on-line information are poorly connected to the network, or very remote from Australia (eg. Finland), and/or are hosted on systems that are not dedicated to providing on-line information (ie. are general purpose computing systems) some users are discouraged from making use of the information services available to them.

Obviously some of these points are contradictory; poor access tools and poor file transfer performance from some sites are likely to reduce the load on AARNet's international line rather than increase it. These issues of user acceptance are genuine however, and should if possible be addressed.

The project proposed in this paper offers a course of action that attempts to resolve the oversubscription of the satellite link to file transfer, and provides an infrastructure upon which improved user access facilities can be developed.

4. Proposal Details

The AARNet Archive Working Group proposal calls for the acquisition, installation and configuration of a computer system to provide an archive of electronic information resources to serve all AARNet members. The system would be configured with a large amount of disk space and would support the following functions:

1) Mirroring

By maintaining exact and up to date copies (mirrors) of some of the more popular overseas information sources, AARNet users will be able to conveniently, quickly and efficiently retrieve the items they require. Most of the maintenance required to update the copy of the source archive is carried out during the night, thus further distributing the network load throughout the entire day;

2) Caching

To avoid the rush of duplicate transfer requests that the announcement of new or updated programs, documents or data generates, the AARNet archive system will automatically retrieve a copy of any document announced in the forum that exists for this purpose (the news group comp.archives). Users reading comp.archives are then assured of finding the item they require locally;

3) Information Resource Identification and Location

The AARNet archive system will host several end user information services. The most significant of these will be an on-line facility that allows users to locate the source of information items throughout the world (but naturally will refer to local copies when available!). Arrangements have been made with McGill University in Canada to host their "archie" package to service this requirement. Remote access to similar information will be provided by the directory server developed at the University of Sydney based on the "finger" protocol;

It is also hoped that by gathering many commonly used items of information together, more inexperienced users will find it easier to locate and retrieve documents and software that is relevant to their work, thereby increasing the use of the network and the individuals productivity.

4) Information Distribution

As well as providing the ubiquitous access method, anonymous ftp, the archive system will support the retrieval of documents via electronic mail and, in the near future, a "fetchfile" facility based on the work funded by AARNet in the 1990 projects.

A single, large, centrally funded and operated archive server has several advantages over the multiple, smaller, distributed systems currently operated by individual institutions;

a) Single point of access.

Users do not have to know which system stores what. It will be found on the AARNet archive server.

b) Improved performance.

Most institutional based information servers are operated on general purpose systems, and thus are often slow to respond to file transfer requests. This problem is exacerbated when the site is not part of the AARNet backbone, and the data being transferred must compete on mid-speed link with the network traffic generated by the AARNet member as a whole.

c) Consistency.

The way data is structured on individual information servers varies depending on the whims of the individual system managers. This complicates the process of locating and retrieving information, but is not an issue on a single system.

d) Economy of scale.

For an information archive, the primary hardware requirement is disk space. By concentrating resources on a single system the overheads of the processor, backup device and memory can be reduced in favour of acquiring additional disk space.

e) Ongoing support.

There is no obligation on the part of individual AARNet members to continue to provide the archive services they currently support, or to archive information of interest to the AARNet community at large. Given the important roles that the AARNet archive server can fill, it needs to be funded centrally, and supported as part of the network infrastructure through the existing hub management agreements.

f) Optimal network usage.

If multiple archive servers were to be deployed, then much of the "maintenance" traffic used to support these servers would potentially travel over longer network paths. A single server, located near the termination point of the international link on the AARNet backbone reduces this usage to a minimum.

The project group therefore recommends that the AARNet archive server be located at the University of Melbourne on the AARNet interconnection Ethernet that links the Victorian, national and international routers.

To determine the required capacity of the archive system, the working group will undertake a survey of existing usage. This will involve the analysis of international file transfer traffic to determine which remote sites are the source of the most of the information volume, and a survey of existing local caching, mirroring and information services sites in Australia.

In negotiating with hardware suppliers for the supply of equipment, the usual parameters of price, performance and reliability will obviously be considered. Other factors that also need to be addressed in this particular case include the availability and cost of disk and memory upgrades, and the extent to which the system software is compatible with existing implementations of the applications that are to be installed.

These applications are listed below, and will be installed by members of the working group, with the support of funding from this project;

	Source	Supports
FTP Daemon	Various	Anonymous FTP
Archie	McGill University	On-line information searches
Finger Daemon	University of Sydney	Remote information searches
Sendmail	Various	FTP-Mail gateway
MHSnet	MHS/AVCC Agreement	Fetchfile gateway

The goal of this first phase of the archive project is to announce a fully functioning service to the AARNet community at the 1991 Networkshop to be held in early December, 1991. This will provide a single initial point of access to all AARNet users for the majority of information access queries. It is hoped that this will reduce the load on the international satellite service, prolonging the life of this service in its current 256K configuration, and improve the response time for international interactive sessions.

The costs associated with this first phase of activity are estimated to be $55,000.

5. Future

A number of programmes are taking place overseas aimed at exactly the issues addressed by this project, ie. access to widely distributed information databases, and on-line information services.

The working group sees the archive project (with appropriate hardware, software and human resources) as an opportunity to participate in the implementation and development of these facilities. In particular the following projects show some promise and are worthy of further investigation, development and funding as part of stage two of the AARNet archive project.

Prospero is a "virtual file system" developed at the University of Washington that allows information stored on geographically separate computing systems to appear as a single store of information. Many of the functions that the AARNet archive server will support (such as caching) are to be integrated into future versions of Prospero with the support of a number of institutions including the University of Wollongong.

In addition to "enabling" technologies such as Prospero, considerable effort has also been concentrated within the international networking community on the provision of end-user information access tools. One example is the "finger" server developed at the University of Sydney which currently provides access to a variety of information sources held at the University of Sydney, eg. internal phone numbers, AARNet resource guide entries. The development of clients for Macintosh and PC systems would make this service more widely available.

On a more global scale, the Wide Area Information Services (WAIS) project has developed information servers and an extremely powerful, very user friendly end-user interface to search, on a keyword basis, through these servers to find documents, maps, images, programs, etc. archived throughout the world. Support for WAIS on the AARNet archive server would be both logical and desirable, as it integrates the search techniques of the existing tools with an improved method of information retrieval.

It is proposed that an additional $15,000 be set aside for this second phase to further the development of archive services, both for the AARNet archive server itself and for improved end-user tools.

At this stage it is anticipated that this activity will be undertaken in the early part of 1992.

6. Project Costings

The selection of an appropriate vendor and hardware configuration to support the archive server will be one of the early tasks to be undertaken by the project working group. However, based on experience and an estimate of the amount of disk and memory (for example) that are required, the project budget shown below budget is proposed:

Hardware (indicative)		
Processor (equivalent to a Sun SPARCserver 2),		
Memory (Total 48-64Mb),		
Backup Device,		
Disk (notionally 5 Gbyte)		
		$52,000
Personnel		
First Meeting (held 12th June, 1991)	$1,500	
Second Meeting (November 1991)	$1,500	
		$3,000
Further Development (end user tools, Prospero enhancements)		$15,000
TOTAL		$70,000

7. Project Management

The AARNet archive server project will be project managed by Peter Elford, AARNet. He will ensure that the working group achieves the goals listed below and document the progress of the project.

The individual members of the working group have existing expertise in specific areas related to the project and will naturally be responsible for those aspects of the project as follows:

Caching, Prospero trials	Stephen Cliffe
Mirroring	Mark Prior
Archie	Craig Warren
Finger server	Ross Cartlidge
FTPMail, MHSnet	Stephen Cliffe/Michael Henry
System Manager	Michael Henry

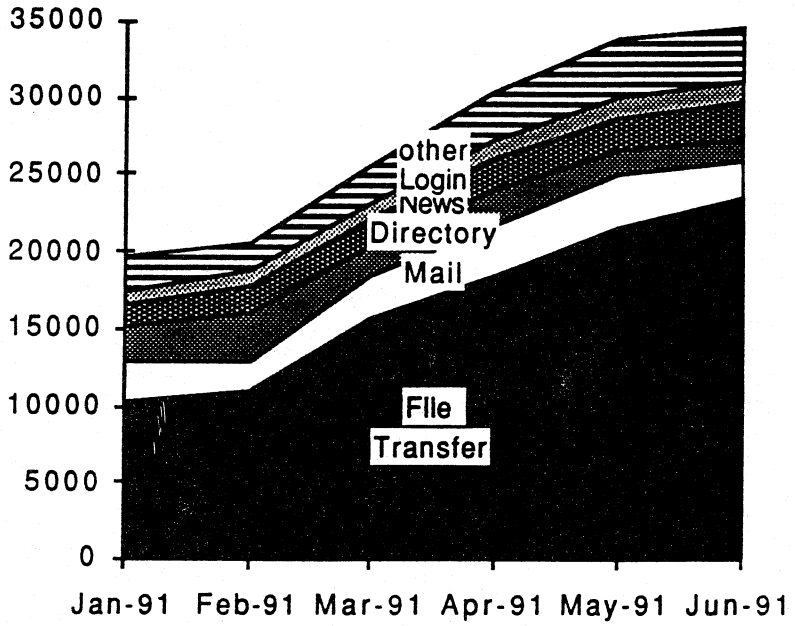
Facilities support will be provided by the University of Melbourne under the auspices of their AARNet hub site agreement.

8. Project Schedule

	Target Completion Date
First Meeting	12th June
Preparation of refined proposal (this document)	mid July
Survey of existing Australia archives, sizing of hardware	early August
Negotiation with suppliers and acquistion of hardware	early September
Second Meeting	early November
Installation and Configuration of software	mid November
Announcement and Commencement of service	Networkshop 91 (early December)
Commencement of Phase 2	January 1992

9. International Link Application Profile

Mega Bytes per month to and from Australia:



Mega Packets per month to and from Australia: