Astronomical Data Analysis Software and Systems VI
ASP Conference Series, Vol. 125, 1997
Gareth Hunt and H. E. Payne, eds.

Automatic Mirroring of the IRAF FTP and WWW Archives

Mike Fitzpatrick and Doug Tody

IRAF Group,[1] NOAO,[2] PO Box 26732, Tucson, AZ 85726

David L. Terrett

Rutherford Appleton Laboratory, Chilton, Didcot, Oxfordshire OX11 0QX, United Kingdom

Abstract. Large FTP archives have long used mirrors (copies of the network archive maintained on remote hosts) to decrease the load on a particular server or shorten the network path to provide faster download times. Little has been done however to simplify mirroring of WWW (World Wide Web) pages, although many projects and users now rely on Web pages at least as heavily as anonymous FTP services. With the dramatically increasing use of the global Internet in the past year, the network has become overloaded, and network access, especially overseas, is often very slow during peak hours. We present a strategy based on host-independent URLs which allows Web pages to be automatically mirrored to both remote Web hosts and CD-ROMs. Issues affecting a site wishing to mirror a remote archive are discussed.

1. Introduction

The subject of automatic mirroring can be approached in one of two ways: from the standpoint of those wishing to export their archive for mirroring, and of those wishing to be a mirror for an existing archive. Although this paper deals with the specific issues we faced in setting up a mirror of the IRAF archives, the techniques presented are general, and can easily be applied to any similar archive.

On both ends there were some expected setup glitches in trying to verify the thousands of links involved, in bringing both systems to a common understanding about requirements in HTTP server and local software, and in establishing a routine procedure for maintaining the mirror. The initial experiment between the NOAO IRAF Group and the UK Mirror at Rutherford Appleton Labs has worked out many of these problems, and has provided us with the ability to establish other mirrors much more easily. In the first five months of operation, the

[1] Image Reduction and Analysis Facility, distributed by the National Optical Astronomy Observatories

[2] National Optical Astronomy Observatories, operated by the Association of Universities for Research in Astronomy, Inc. (AURA) under cooperative agreement with the National Science Foundation

UK IRAF Mirror Site has distributed more than 4300 files to 120 different nodes in more that ten countries, providing a faster, more reliable link for UK and European sites. Negotiations are underway to establish mirrors in other parts of the world where FTP access to the NOAO Tucson archives or UK mirror sites is prohibitively slow.

The host-independent manner in which the WWW pages are written means that they can also be used from a CD-ROM running on a local machine, in effect duplicating the IRAF archive on any machine. We discuss the limitations and special setup required in this case.

2. Preparing Your Archive for Mirroring

There are only a few steps involved in preparing your archive so it can be easily mirrored elsewhere:

2.1. Host-Independent URLs

The mirroring site will have a Web address different from the original site. If Web pages contain explicit HTTP URLs, then these pages will still refer to the *original* archive when the pages are mirrored, negating the point of the Web mirror. The simplest solution is to substitute file relative URLs in all cases except where one really does want a URL to point to a specific network host. For the exporting site this means each link will need to be examined and changed in the following ways:

- Use "file.html" or "subdir/file.html" link references. Keep it simple, no complex relative paths.
- Since the Web root directory may be different on the mirror node (which is likely serving its own documents), root-relative links such as "/iraf-homepage.html" should be avoided.
- We don't want to require that the mirroring site put documents in a particular directory, so the best compromise is to establish a set of common links for both systems so root-relative paths can be used on either host correctly. In our case we established /iraf/ftp and /iraf/web links pointing to the root of the FTP and WWW areas respectively (this also fits in well with the suggested directory structure for an IRAF installation). It so happens that our Web pages are under the FTP directory tree, but this is not a requirement.

2.2. CGI Scripting

There are several things to be done to make most CGI scripts portable:

- One cannot assume that a mirror node will have all of the custom local software that may be used by CGI scripts, or indeed that it is even the same type of machine. For our Web archive we've created a *bin*, *lib*, and *src* subdirectory containing binaries and source for all programs (mail filters, search engines, etc.) used by the various scripts. All binaries are built from these sources, meaning that versions are current for all platforms and are automatically updated in the mirror site when a new version is installed by the originating site.

- In HTML forms or links, references will be made to a particular script or application. Since a mirror may be running on a type of computer different from the original server, these task names are actually csh scripts which call the binary (in the *bin* subdirectory) appropriate for that platform. In our case, the scripts reference a program called *mget* as a mail filter, the *mget* script figures out what type of machine it's running on and passes the arguments to a *mget.sparc* binary to do the actual work.
- Path names in scripts CGI scripts are often written as scripts of some type (csh, Perl, etc.) which are invoked using an, e.g., `#!/bin/csh` path as the first line. Such paths may not be universal, however. The mirroring site is responsible for creating the system links needed to resolve these paths.

2.3. The Final Steps

You may wish to arrange for mirror site usage logs to be propagated back to the original site. This can be done as a weekly cron job that greps for entries containing a certain keyword in the logs ("iraf" in our case) and automatically mails them to a specific maildrop. If the archive is large, it is best to make a snapshot tape of the full directory tree tree to be mirrored and mail that to the mirroring institution to populate the initial directory tree. Once the initial system is installed and working, updates should be small and will be handled automatically by the mirror software.

3. Setting up a Mirror Archive

Now that the initial IRAF mirror site has been established, we should have worked out most of the bugs in the scripts and documents on our end, but there are still concerns for new sites wishing to establish a mirror:

3.1. Disk Space Required

The complete IRAF archive now requires approximately 3 GB of storage—this will probably increase another 1 GB in the next year as more software is released. Potential mirrors should consider the purchase of a new dedicated disk.

3.2. Mirroring Package Used

The RAL mirror site is maintained using a package called *MIRROR* from Lee McLoughlin of the University of London; other packages are also available. This particular package required Perl 4, which had to be installed specifically to support the package. A cron script is run nightly to update the archive, and a separate script is run once weekly to mail access logs back to Tucson. The archive scripts directory is mirrored separately to a different directory, in part because execute permissions are stripped in the mirroring process and in part so new code may be hand checked, as a security measure.

3.3. HTTP Server Requirements

The UK mirror was already serving Web documents and had a configured HTTP server. New sites, or those using the CD-ROM, may need to configure a server. The only changes required to support the mirror were alias definitions for the IRAF CGI scripts directory. This means editing the httpd/conf/srm.conf file with an *Alias* and *ScriptAlias* definition for the scripts directory which points to the iraf Web scripts directory on the mirror, and aliases for the root-relative links. For example,

Alias	/iraf/web	/iraf/web
Alias	/iraf/ftp	/iraf/ftp
Alias	/scripts	/mirror/iraf/web/scripts/
ScriptAlias	/scripts/	/iraf/web/scripts/

One other problem is that most HTTP servers define a default MIME type as plain text for documents for which the server cannot determine the type from the file name extension. This means that tar files, compressed PostScript files, etc., show up as jumbled text in the browser rather than being identified as binary or starting an external viewer. To work around this, we suggest the following definition in the server's srm.conf file

```
Redirect     /iraf/ftp/     ftp://iraf.noao.edu/
```

This causes the most browsers to create a save pop-up window rather than trying to display the file, which is what is most often desired.

Aside from the initial setup and verification of new scripts, the process is now largely automatic requiring an estimated one hour/month to maintain the mirror. Only rarely has the nightly update not completed successfully; in each case it has succeeded the following night.

4. CD-ROM Issues

While the host-independent nature of the WWW pages means the archive can be distributed on and browsed from a CD-ROM, there are a few issues of concern for viewing the CD-ROM Web pages as though they were a *live* Web site:

- An HTTP server must be running in order for CGI scripting to work, otherwise all documents are accessed with a file URL and scripts will not execute.
- Sites using the CD-ROM as a local archive must also remember to set the /iraf/ftp and /iraf/web links discussed, so links are resolved.
- At present, the archive only contains binaries used by CGI scripts for those mirror platforms we know about. More binaries are needed.

5. Project Status

We welcome inquiries from any sites wishing to set up additional IRAF mirrors, or from sites interested in using the techniques outlined in this paper to mirror their own archives. Contact *iraf@noao.edu* for further information.