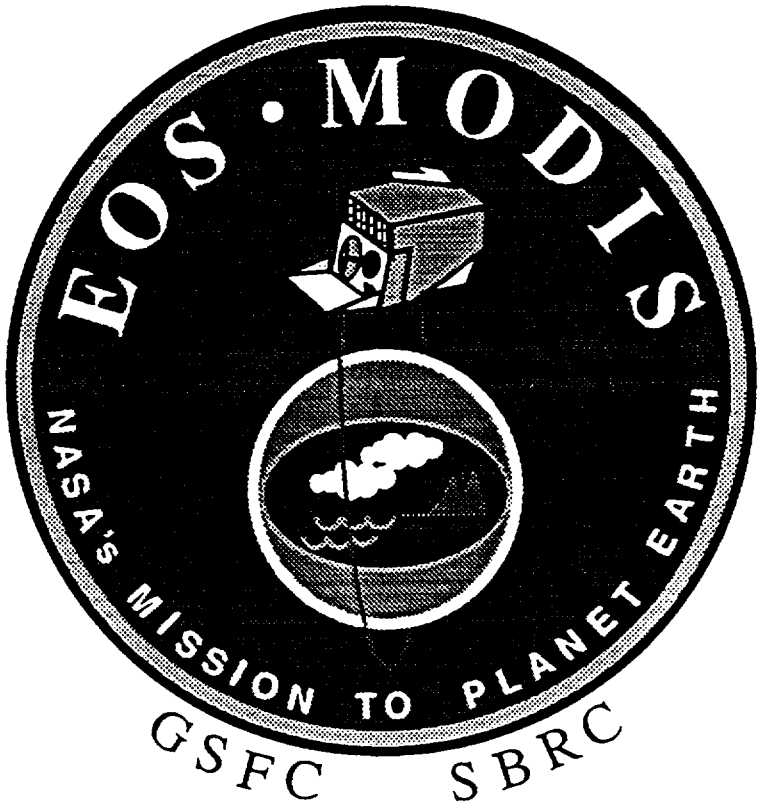
MODARCH Presentation
at the
MODIS Science Team Meeting

September 30, 1993



by

David Herring & Michael Heney

Science Systems & Applications, Inc.

1.0 INTRODUCTION

1.1 Statement of Problem

Late in the summer of '92, Locke Stuart (former MAST Team Leader) directed Paul Baker (Presidential Management Intern) and David Herring (MAST Technical Manager) to answer a fundamental question: are our current archiving methods sufficient, or does MODIS need an electronic document archive? Additionally, if we decide we need an electronic archive, what features should it have and what's the most efficient and cost effective way of setting it up?

1.2 Research Findings

Baker and Herring then designed a study to determine the Science, Technical, and Support Teams' requirements for the MODIS archive. The objectives of the study were to:

- review MODIS' and other related document archives to see how other organizations are setting up their systems;
- survey team members to determine their needs and preferences for reading, writing, and transferring documents;
- investigate archiving technologies;
- scope the size of the archive and the number of users;
- develop system architecture and cost options; and
- recommend the system which best meets the team's needs.

1.2.1 Review and Scope of MODIS' Current Archive

Baker and Herring conducted interviews and found that there are a number of critical problems inherent in the existing archive strategy. One of the foremost problems is space—currently the archive consists of about 63,000 pages which fills almost 6 filing cabinets. At a 20:1 compression ratio, this equals almost 6 Gb of electronic storage space. They found that the rate of document input into the archive is increasing and estimate that 10 years from now the archive will require 85 filing cabinets, or about 85 Gb for electronic storage.

As the number of filing cabinets increases, so does the difficulty in retrieving information quickly and precisely. Moreover, as space is somewhat limited at Goddard, 85 filing cabinets would be an inefficient consumption of space. Searching for specific information among 85 filing cabinets would be an inefficient consumption of time.

1.2.2 The Needs of MODIS Archive Users

Obviously, switching to an electronic system must improve upon the existing system—it must enable users to retrieve information more quickly and with a higher degree of precision. (It is a given that any electronic system will consume a tiny fraction of the space filled by the current system.) An electronic MODIS archive, by definition, has to be accessible to the entire team with a minimum amount of effort. It must be extremely user friendly, which means users must be able to navigate through it intuitively from their respective desktop computers. Therefore, the system will have to be flexible because the Science and Support Teams do not all use the same computer platforms. Everyone on the team uses one or more of the following computer platforms: Macintosh, PC, and UNIX. Subsequently, so as not to exclude anyone on the team, Baker and Herring restricted their review of electronic archiving technologies to only those compatible with each of these three platforms.

Additionally, because most of the archived documents are hardcopies, the system must have a means for

converting the documents to electronic form. Therefore, the archive must have a scanner and optical character recognition (OCR) software. However, no OCR software is foolproof—it is estimated that the best packages still misread 1-5 characters out of every 100, even on good-quality laser-printed copy. This error rate is unacceptable; if there are an average of 250 words per letter-sized page and 1-5 out of every 100 characters are misread, then there could be as many as 100 misspelled words per page in a scanned document. Hence, the system must have a way to correct or allow for OCR errors if a retrieval algorithm is to be run with an acceptable degree of precision.

In summary, the ideal system should allow for input of both hardcopy and electronic documents into a computer archive which contains mass storage media, OCR software, indexing software, compression software, and software for searching for and retrieving information. It should also be accessible from multiple platforms (Mac, PC, and UNIX) for output in either electronic or hardcopy form.

1.2.3 Recommendations

Based on the needs and preferences of the MODIS Team, and their review of the software, Baker and Herring recommended procuring PixTex/EFS. The decision was made to purchase the software through INTRAFED, as it comes bundled with the necessary hardware components in a single, integrated system. Their system—the "stand-alone, single-user imaging system"—was selected because it is an almost "turn-key" solution.

On June 30, 1993, the system was procured. In order to assign it an ethernet address, the system was named "MODARCH" (for MODIS Document Archive).

On July 7, 1993, Michael Heney reported to work as MODARCH System Administrator.

1.3 Summary of MODARCH Features

- MODARCH will allow users to retrieve information at their desks in seconds with the intuitive, point-and-click ease of a Macintosh
- MODARCH will consume a tiny fraction of the space which would be consumed by filing cabinets
- MODARCH allows access from Macs, PCs, and UNIX computers
- MODARCH requires minimal effort for submission of documentation
- MODARCH allows for multiple search and retrieval strategies, including a fuzzy search capability
- MODARCH eliminates the need for distribution of hardcopies
- MODARCH provides allows everyone access to one common, central database

2.0 MODARCH CLIENT SOFTWARE FOR MACINTOSH

Excalibur's Pix/Tex EFS Client software for Macintosh systems is available via anonymous ftp from the MODARCH system. In order to run the client software, your Macintosh needs to be running System 7.0 or greater, and MacTCP must be installed.

In brief, the information that you need to get the client software is:

hostname: modarch.gsfc.nasa.gov (128.183.26.44)
ftp directory: pub/EFS-MAC
filename: EFSv3.0.1-tcp

The following example ftp session was run using NCSA Telnet. This package is available via AppleShare at GSFC. From the Chooser, highlight AppleShare and select !Mac Network Support as the AppleTalk Zone. Select 251.6 Mac Network Info Ctr as your file server. Connect as Guest, and select Goodies from the menu "Select the items you want to use". Goodies will appear as an AppleShare item on your desktop. Open it, open the folder TCP-IP Software, and from there open the folder NCSA Telnet. Copy the folder NCSA Telnet 2.5 to your desktop. You will use the application NCSA/BYU Telnet 2.5 for your ftp session.

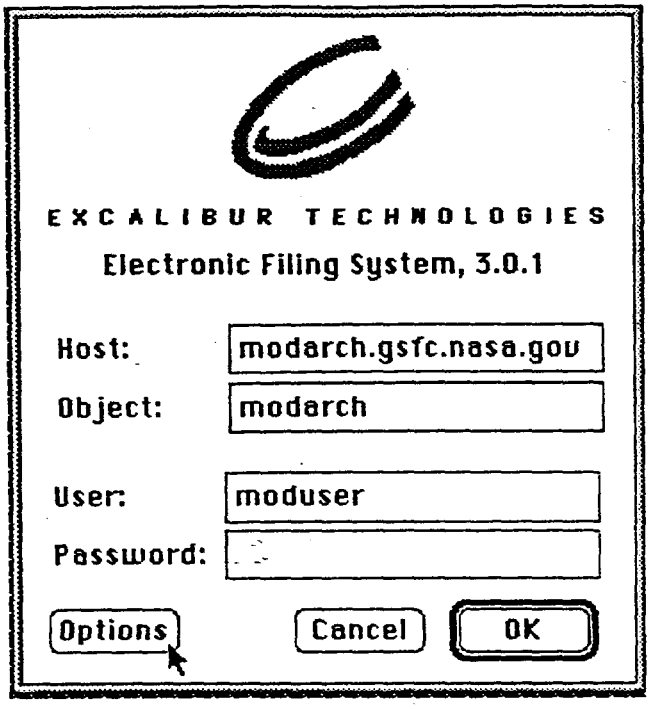
When you start NCSA Telnet, go to the File menu and make sure the FTP Enabled and MacBinary Enabled options are both selected. The **MacBinary option is especially important**, as it tells your Macintosh to treat the downloaded file as an application rather than a binary data file. Also, you should set your transfer directory (File menu option) before issuing the mget command. If you are using something other than NCSA Telnet for your ftp connection, make sure that the data transfer options are configured properly to download the file as an application.

To start an FTP session, go to the File menu and click Open Connection. In the dialog box that appears, select FTP Session. In the box labeled Session name, enter the hostname - just modarch will do if you are on-site at GSFC, otherwise enter the full name modarch.gsfc.nasa.gov. Click OK, and the session begins.

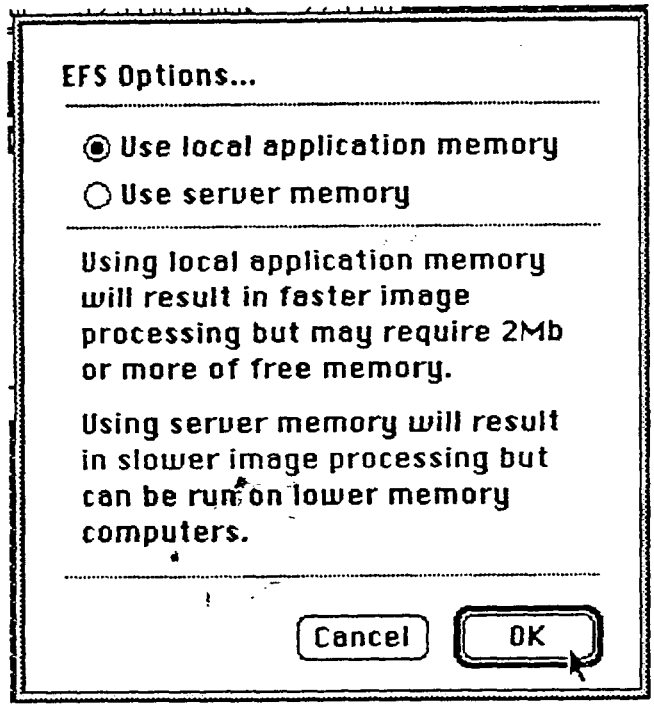
Here is a sample FTP session to get the Mac EFS Client software. User input is in *italics.*

```
220 modarch FTP server (SunOS 4.1) ready.
user ftp
331 Guest login ok, send ident as password.
yourname@host
230 Guest login ok, access restrictions apply.
cd pub/EFS-MAC
250 CWD command successful.
bin
200 Type set to I.
mget EFS*
200 PORT command successful.
150 Binary data connection for /bin/ls (128.183.26.141,44896) (0 bytes).
226 Binary Transfer complete.
Receiving EFSv3.0.1-tcp
200 PORT command successful.
150 Binary data connection for EFSv3.0.1-tcp (128.183.26.141,44895) (389888 bytes).
226 Binary Transfer complete.
bye
```

3.0 FIRST LOGGING INTO MODARCH



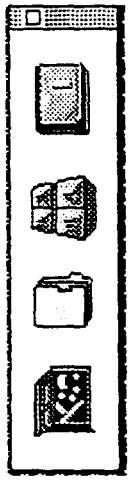
Double-click on the EFS icon on your desktop—the window pictured on the left will appear. Enter the text as shown, being sure to leave the password field blank—you will not need a password to access MODARCH. You need only enter this information once, EFS will remember after that. Hit the "Options" button and the window pictured below will appear.



3.1 Use Local Application Memory

Select "Use local application memory" if you have 4 MB of RAM, or more. If you have less than 4 MB of RAM, leave it set at the default setting of "Use server memory". Click "OK" on this and the previous window to begin using MODARCH.

3.2 EFS Tools Palette

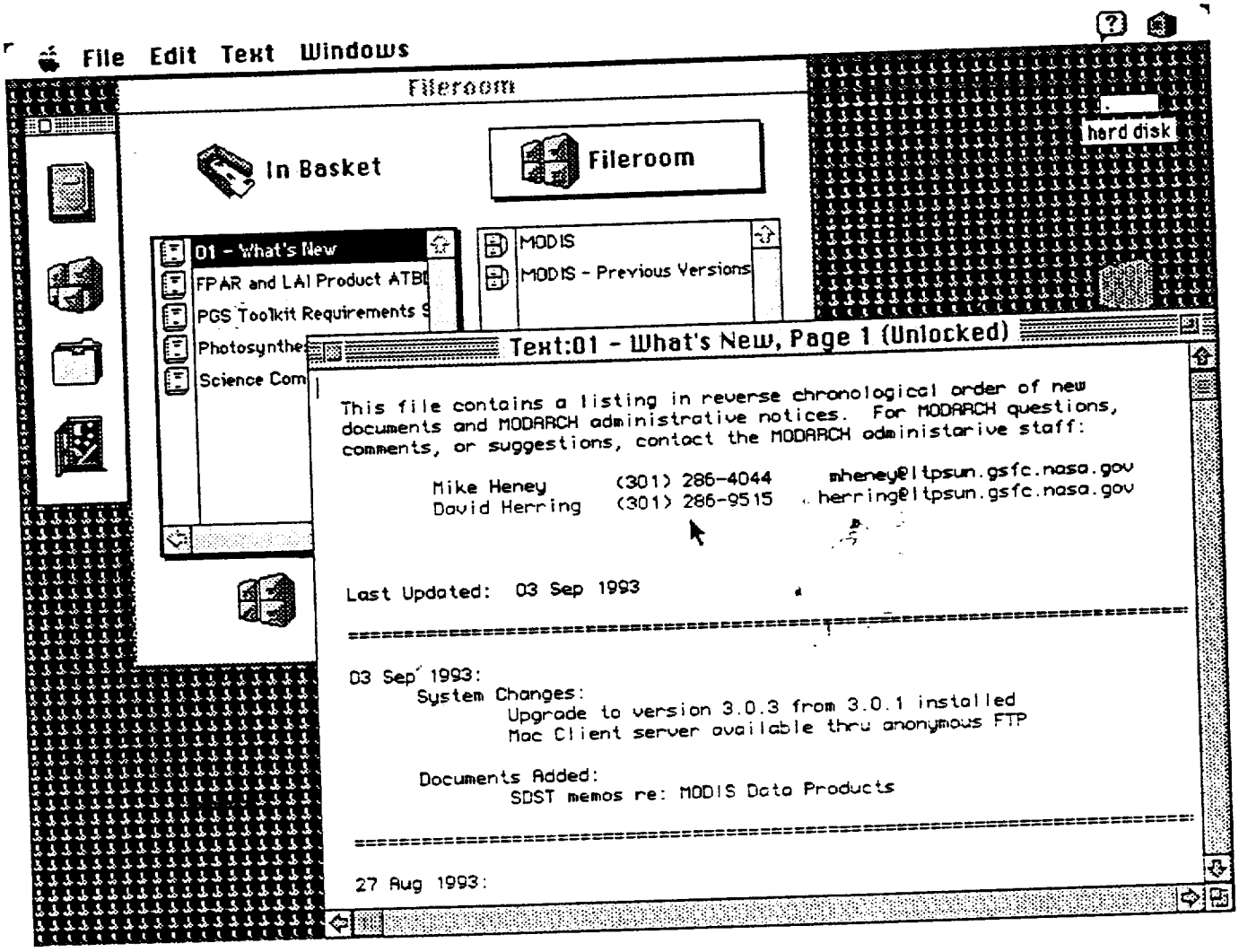


MODARCH is organized like a conventional fileroom to enable you to intuitively search for and retrieve information. By pointing and clicking on familiar icons—filing cabinets, drawers, folders, and documents—you may proceed directly to and view documents.

Initially, upon accessing MODARCH, the EFS Tools Palette appears in the upper lefthand corner of your screen. Use the Tools Palette to navigate to other EFS windows. By clicking on one of three icons in this palette you can access windows where you can search for information within MODARCH. Click on the Document (top) icon to open the Document Window; the Fileroom (second) icon opens the Fileroom Window; the Folder (third) icon opens the Search Window; and clicking the Exit (fourth) icon exits MODARCH.

3.3 What's New

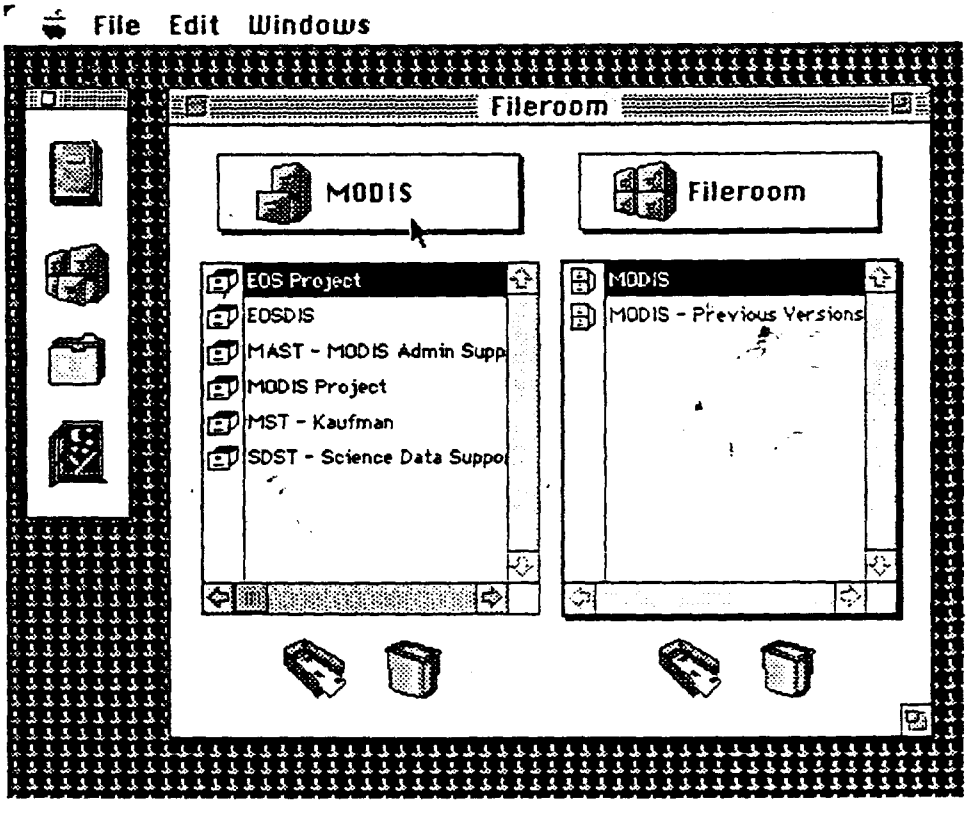
Whenever you log onto MODARCH, review the "What's New" document in the In Basket window. This document contains a listing in reverse chronological order of any new documents or MODARCH administrative notices entered into the archive each week. "What's New" will be updated every week to provide the MODIS Team notification of release of important documents. "What's New" will also indicate where the document is located in MODARCH.



4.0 DOCUMENT SEARCHES

Once you become familiar with the Fileroom organization, the quickest means of retrieving information will be to proceed directly to the specific cabinet, drawer, and folder containing the document(s) you desire. To enter the Fileroom, simply click the Fileroom Icon. There are currently two filing cabinets in the Fileroom—"MODIS" and "MODIS - Previous Versions". The "MODIS - Previous Versions" cabinet will be organized exactly like the "MODIS" cabinet; however, it will contain only older versions of documents.

The EOS Fileroom will be oganized according to the flowchart in Figure 1 (see next page). The window below shows the drawers currently comprising the MODIS cabinet. Most documents are placed in drawers according to the person or group or organization that created them. For example, if you want to view the MODIS Specifications, double-click on the "MODIS Project" drawer.



4.1 Content Searches

Clicking on the Folder Icon (third from the top) opens the Search window, enabling you to search for and retrieve *documents*. The success of your search depends upon the relevance of your clue, which may be up to 128 characters long. There are three methods by which you may search for a document—Content, Label, or Control searches. You will use different strategies based on what you already know about the document you wish to retrieve.

Click in the Clue field and type "volcano". Clicking the "Content" button tells MODARCH to find "volcano" in the contents of every document in the EOS Fileroom. Use content searches when you

MODARCH Organization

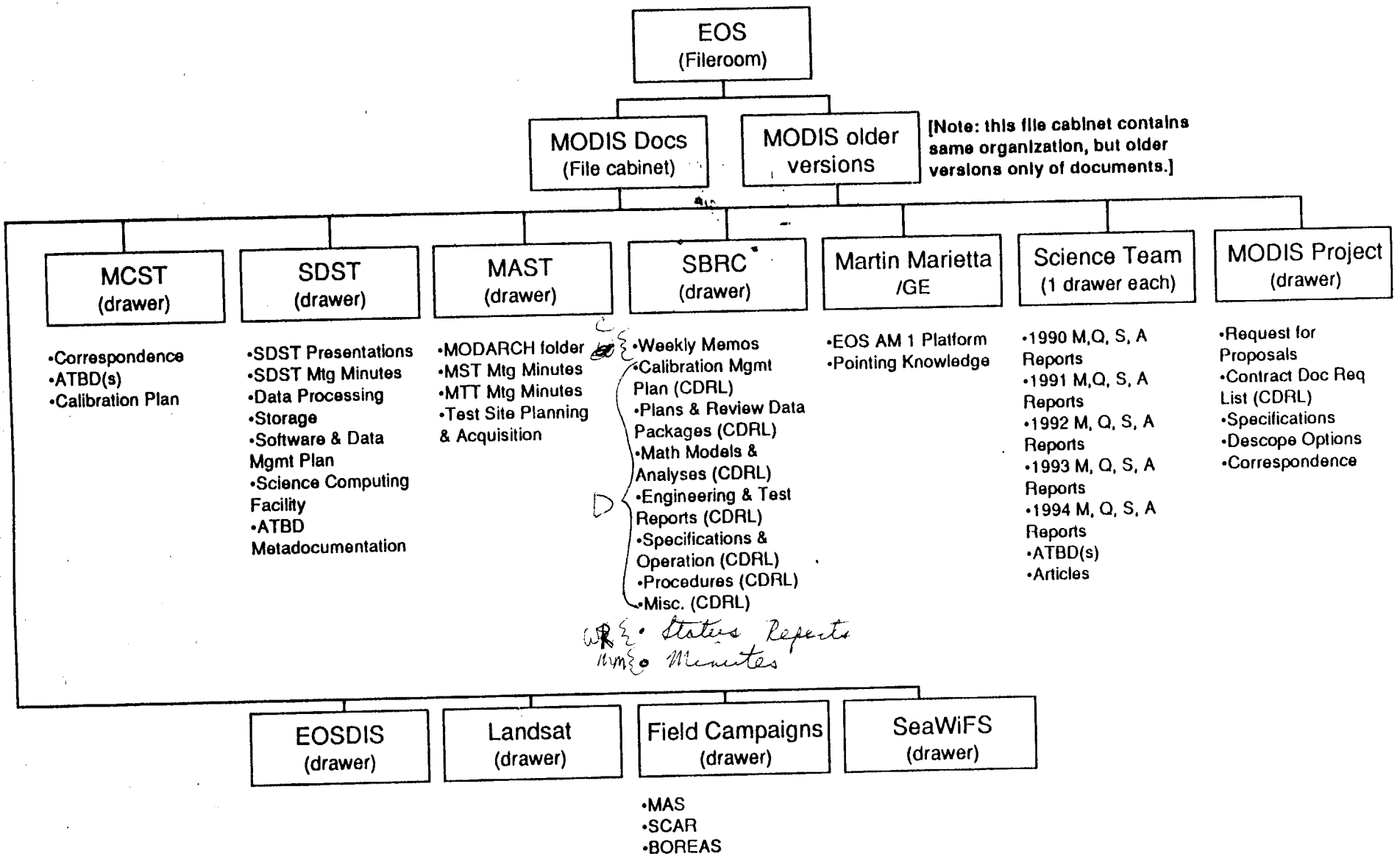
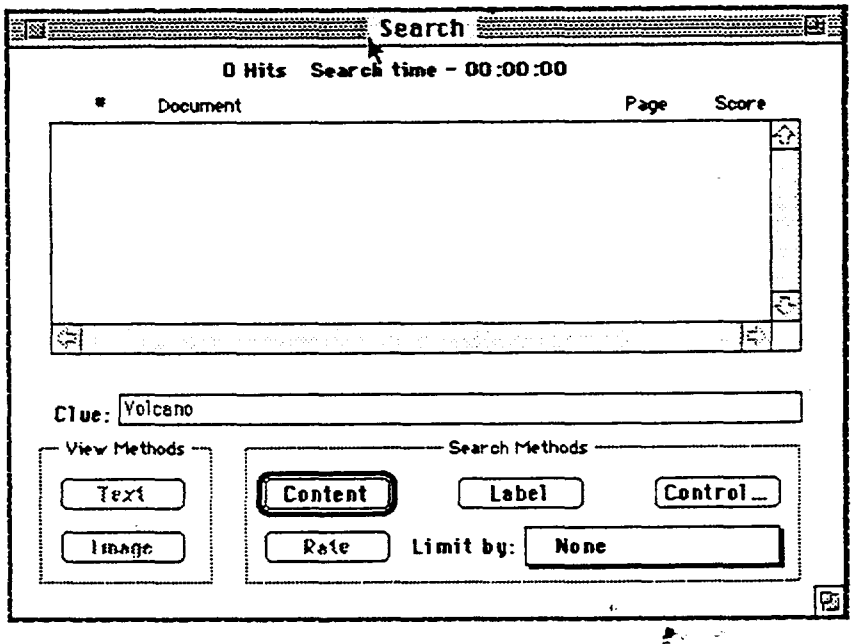
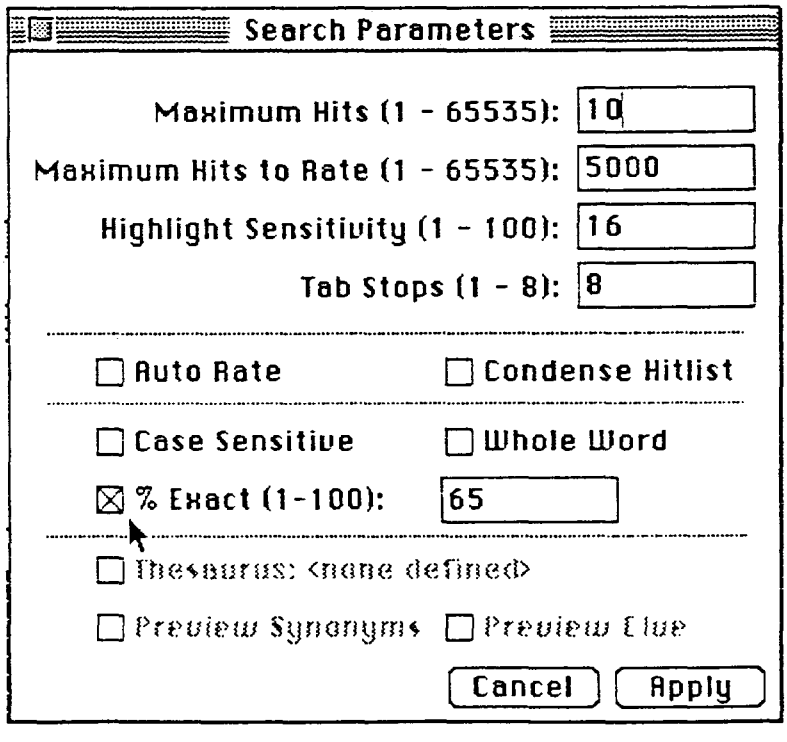


FIGURE 1

know what information you want, but you don't know where it is. Please note, unless indicate otherwise, MODARCH will do a *fuzzy* search which means that it will not only find "volcano", it will also find variations of the word as well as similar words. You may either do an exact or a fuzzy content search. Fuzzy content searches are useful if you don't know how to spell what you're looking for, or if it is likely that there are errors that occurred in the Optical Character Recognition (OCR) process when the document was scanned into MODARCH.

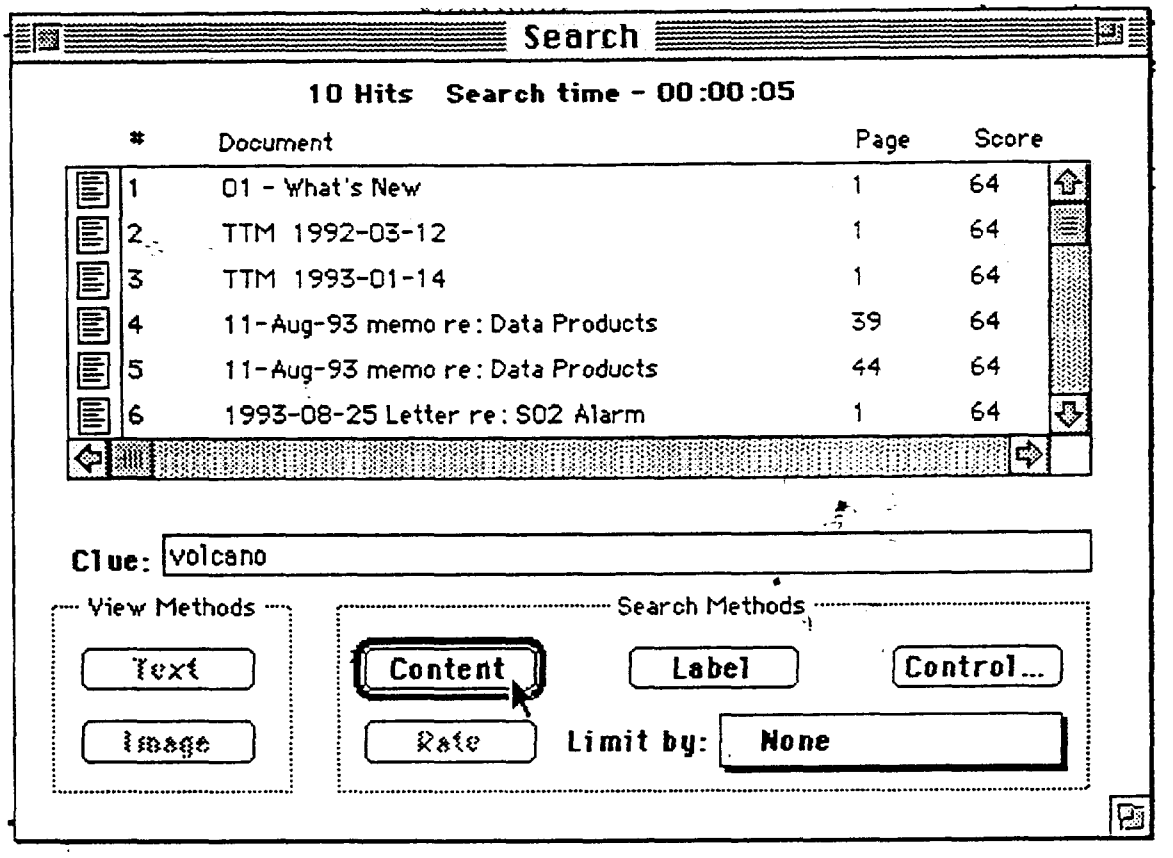


To set the parameters for your search, select **Parameters...** from the **Search** menu. Select the % Exact (1-100): field and then indicate the degree of fuzziness for the search you would like to conduct (the default is 65). Or, if you prefer, you may deselect that field to conduct an exact search.



Upon completing a Content Search, a hit list will appear indicating the hit number, document name, document page number on which a hit was made, and a relevance score. To view one of the hits, simply click on the document name to highlight it and then select either "Text" or "Image" in the lower lefthand corner of the window. *Important note:* Your hit list will consist of *individual pages* contained within documents.

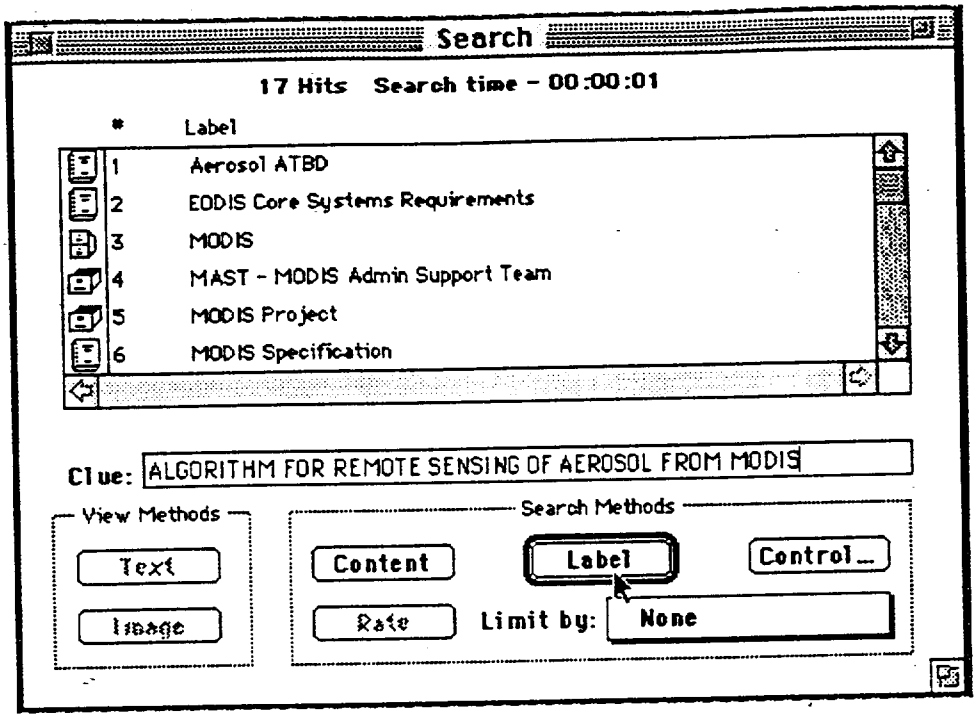
You may limit your search display to a pre-selected object within the fileroom, or to the contents of a previous hit list. To do so, click in the Limit by: field in the lower righthand corner of the Search window, select the desired limit, then click "Content" again.



4.2 Label Searches

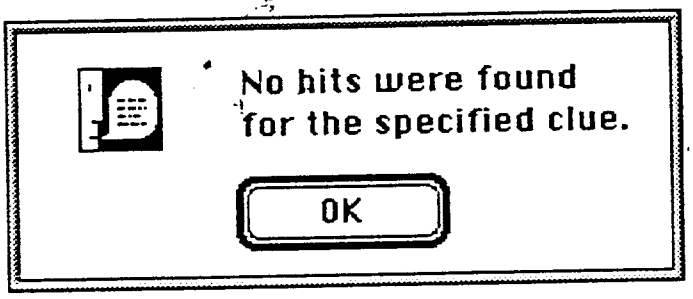
Label Searches find specific documents, folders, drawers, or cabinets in the fileroom according to the object's label (or name). As is the case for Content Searches, you may do an exact or a fuzzy Label Search. You may also limit Label Searches. A Label Search is the easiest method for retrieving a document if you know a document's precise or approximate title. *Important note:* For Label Searches, your hit list will consist of *objects* within the EOS Fileroom—i.e. cabinets, drawers, folders, or documents, but not individual pages.

Document's labels closely correspond to their actual titles. For example, if you know the title of Yoram Kaufman's ATBD (Algorithm Theoretical Basis Document) on aerosol, simply click the Folder Icon and enter the title in the Clue field (as shown below). Next click on "Label". MODARCH found the correct document and gave it top ranking because its label is similar to the document's actual title (in this case it matched the word "aerosol").



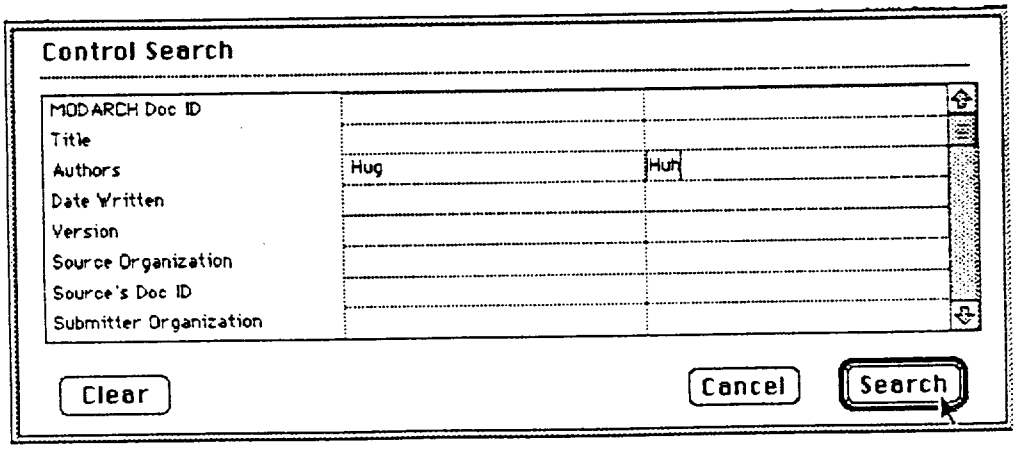
4.3 Control Searches

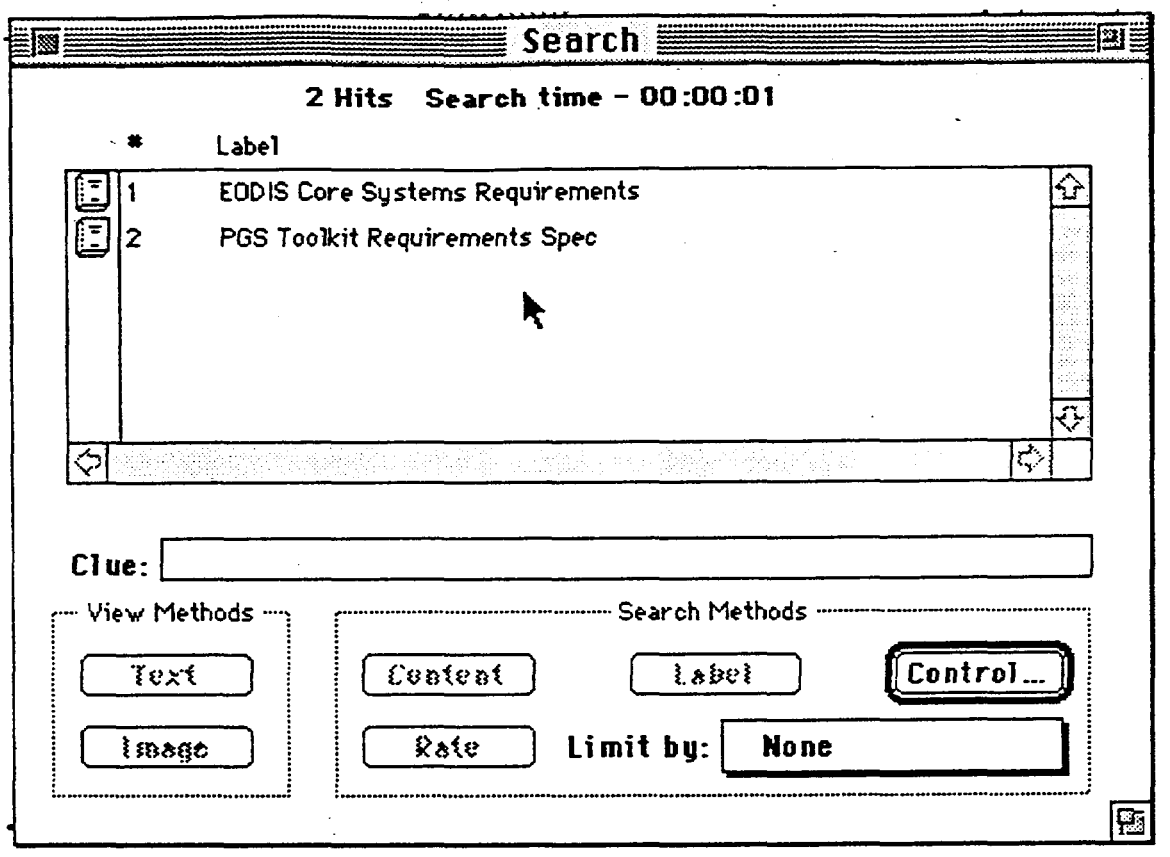
Control Searches allow you to search for documents based on Document Control (or index) information. These searches are exact only; however, you may search for documents that fall within specific ranges. For example, you may search all documents published between the dates 01-JAN-1993 and 14-SEP-1993.



To do a Control Search, click the Folder Icon then click the "Control" button. The Control Search window will appear containing the list of fields according to which every document in MODARCH is archived. For example, if you want to see a list of every document produced by Hughes you would type "Hughes Applied Information Systems, Inc." in the Author field and click "Search". However, because Control Searches are exact searches only, if you misspell a word, enter any incorrect punctuation, or type a letter with an incorrect case your search will be unsuccessful.

To avoid this problem you can either take the time to make sure you enter your clue exactly right, or enter a range. For example, in the Author field enter "Hug" in the first cell and "Huh" in the second and click "Search". This search method allows you flexibility while still generating the desired hit list.





5.0 USER LIMIT

Currently, MODARCH is limited to five concurrent users at one time. The window on the right will appear if you try to log on and the limit is already met. Simply click "OK" and wait awhile before trying to log on again. If this problem persists, contact the MODARCH system administrator (Michael Heney). It is possible users may log on and then not log off once they have finished.