THE FUTURE OF ARCHIVE SERVICES AT SPACE TELESCOPE (FASST)

Megan Donahue (co-chair), Niall Gaffney (co-chair), Stefano Casertano, Harry Ferguson, Bob Hanisch, Ed Hopkins, Cathy Imhoff, Tim Kimball, Mark Dickinson, Chris O'Dea, Daryl Swade, Rick White

NOV 14, 2001

TABLE OF CONTENTS

THE EXECUTIVE SUMMARY

The working group on the Future of Archive Services at Space Telescope (FASST) was a joint committee sponsored by MAST/AB and ESS to review the future of the archive over the next five years. For three months, June-August 2001, we met on a weekly basis to review current services both internal and external to the Institute. We charged the scientists in the group to identify what new services could improve the usability and scientific productivity of researchers using a variety of archives and to assess the scientific value of providing those services. The engineers and software designers were charged with assessing the technology and the technical requirements of such services, in dialogue with the scientists.

What this document is:

- An input to the overall strategic planning for the MAST, to be considered alongside documents from the MAST Users' Committee, SHARE, and MAST's operational goals and objectives, including the on-going replacement of the HST archive engine.
- An assessment and identification of gaps in the service coverage of current archives and suggestions as how MAST might step in to fill those holes.
- An assessment of the scientific priorities of future MAST services.
- A view forward to MAST's role in the NVO.

What this document is not:

- A strategic plan for MAST.
- A specific roadmap to the NVO for MAST.
- A technical implementation plan for future interface projects.

Our top six recommendations, listed in approximate order of priorities, follow on the next page. This table is not a complete list of the FASST recommendations, nor does it contain all of the technical details. In that table, we include a brief description of the scientific value and a technical assessment of difficulty and approximate level of effort for specific implementations of the project. Under the project activity/name, we cross-reference to our outline of project categories with more detailed descriptions on pages 11-14.

Project or Activity	Brief Description	Scientific Value	Technical Assessment	Level of Effort
Data Portal for MAST (1)a)	Enhance the current MAST website for data discovery: enable common, multi-archive searches and retrievals with a general-use astronomical portal, generic catalog, query, and retrieval descriptions.	Allow one-stop shopping for all MAST data.	Technology available. Web and StarView development should be coordinated. Review current MAST website. Leverage existing services and significant domain expertise at STScI.	2 FTE for 1 year, 1 FTE to maintain and enhance; from current StarView and MAST web development; science oversight. 2 weeks of work per domain expert to compile links and services. Standards work ~ 1 FTE for 1 year.
Coordinate MAST catalog services (2)a); coordinate with external catalog services (a catalog browser plug-in, including plotting and statistical services.) (2)b)	Enable users to utilize the information obtained from multiple, independent sources.	Sample construction & data discovery become a lot easier and faster.	Technology is available, but implementation would have to be scoped before committing to this project. One could start with MAST, Sloan, and GSC2 object catalogs.	1 FTE for a year. The browser catalog plug-in, and the plotting plug-in are not large projects but require expertise or training for current developers.
Create standards for contributing user data to MAST. (4)a)	Users should be able to find standards and templates for contributing data and other information to MAST. First demand for this service may come from HST Treasury and archive legacy programs.	Other users will benefit from the processing efforts and the expertise of their colleagues.	2-3 scientists would work with the MAST interface group to create useful templates and standards. Templates should be update, taking into account usage and feedback from users.	0.05 FTE/person for 5-10 people for a year, not including maintenance efforts. Some of this time has already been invested.
Develop transition plans for active to heritage instrument status for HST. (5)	A clear understanding should exist between MAST and STScI instrument groups so that data, documentation, source code and other essential items migrate smoothly from the care of the instrument group to the archive.	Data will maintain its usefulness long after the instrument has retired from active duty.	A small team composed of representatives from the archive and the instrument teams should meet well before an instrument is retired. For currently inactive instruments as well as active instruments, ECF and the CADC should participate in the agreement. Start with the FOC? Learn from FOS experience.	Creating the agreement might need 0.10 FTE/person for 3-4 individuals divided between AB, ESS, and HST over 3 months. Putting the plan into action must be low cost; we estimate ~0.2 FTE for one year before transition for *archive-specific* transition activities.

Improve MAST meta-data: collaborate on the SHARE WCS improvement for HST data: (3)a), (4)b), (4)c). These efforts are grouped because of their similarity to SHARE work.	WCS coordinates need to be improved for other data enhancements, such as stacking, to be reliable. A special effort should be made to correct the worst cases for HST (pointing off by over 1-3"). There are feasibility concerns.	Catalogs and other extracted datasets rely on the accuracy of the header coordinates. Future services such as co-addition will be more reliable if coordinates are reliable.	SHARE also recommends this action; the FASST group defers to the SHARE group for implementation estimates.	The archive may expect to update keyword and database information and to assist instrument scientists in their assessment of the archive contents.
Improve MAST meta-data: add limiting magnitude or surface brightness (3)b).	Go beyond just filter and exposure time combinations that require outside expertise: add a field like limiting magnitude to the data for assessment of usefulness. Other quantities may be extracted with augmentations to reduction software.	Allow non-experts to assess whether data meets their scientific needs.	Expertise is available; one could use the current ETC software to make the computation. Develop a step in the calibration pipeline to populate the header automatically. Backfill old data. Triage: do this for WFPC2 first.	1 FTE for 1-2 months/instrument, from instrument team. Developer/ scientist time split 70/30, including algorithm and pipeline testing. Backfilling a database table with a formula based on fields in that table is not difficult.

The FASST team has presented the basics of this plan to the MAST Users' Group and the Space Telescope Users' Committee. Suggestions and questions from those presentations are taken into account in this document. This white paper is intended for the consideration of the archive branch chief, who leads MAST, the ACDSD division head, and the ESS division head. The FASST recommendations are suggestions intended to prepare the archive for the demands of the data-rich environment of the next five years.

INTRODUCTION

FASST CHARTER AND STRATEGY

The working group "The Future of Archive Services at Space Telescope" was given a broad charter by Marc Postman (AB branch chief) and Stefi Baum (ESS Division head) to review possible directions for the archive over the next five years. The committee was to take into account the "National" Virtual Observatory (NVO) efforts, the Next Generation Space Telescope, legacy Hubble Space Telescope, new space missions, and projects. The group also investigated how we could utilize or combine existing services and new technologies (such as software agents and/or remote process control protocols) to extend current functionality of the archive. In this document, we refer to the FASST team as "we" or "FASST" and to MAST (which includes HST in its harbor) as "MAST" or "the archive".

The working group planned to investigate the following questions:

1. What are the main opportunities on the horizon for the archive?
2. What should the archive do to prepare for those opportunities?
3. Can we identify technologies to be used throughout the archive services at STScI that will allow its separate services to work together?
4. What is missing from the suite of services now available to astronomers?
5. Are there duplicate services in-house that can be combined?
6. How can MAST best leverage the STScI in-house resources and experience?
7. What should our priorities be?

The goal of the FASST working group was to define the "big picture" goals for future archive development and the actions MAST needs to take over the next few years to achieve those goals. An additional goal was to suggest policies that will allow the archive to respond quickly in the face of technological advances. FASST aimed to recommend common technologies that will increase interoperability among archive services both inside and outside MAST. The FASST discussions and recommendations are intended to provide the archive both a guideline regarding scientifically motivated extensions to its current services as well as a better understanding of the needs of a National Virtual Observatory (NVO).

We began by defining the high-level goals of all archive services as guided by a few science cases crafted by members. We then catalogued and reviewed the services that are now available to the community. We identified potential services that are not currently available, particularly those that would increase the number and scope of possible science investigations and that would improve the speed and reliability of the research process. We then investigated various means of providing such services. Those services were then ranked independently for scientific utility and for tractability or ease of implementation. We provide the rankings of the services and a recommendation for implementation of these services, including estimates of effort and timescales.

This group is distinct from the parallel SHARE process, which focused on scientific products from the reprocessing pipeline. The SHARE group had two members in common with the FASST group, which helped to minimize duplication of effort.

The FASST membership: Megan Donahue (co-chair, also a member of SHARE), Niall Gaffney (co-chair), Stefano Casertano, Harry Ferguson, Bob Hanisch, Ed Hopkins, Cathy Imhoff, Tim Kimball, Mark Dickinson (SHARE), Chris O'Dea, Rick White, Daryl Swade.

FASST charter and minutes are stored at http://corsair.stsci.edu:6699/fasst/. An informal index of archive services is also posted there. We had seven meetings; one week was allocated to investigating browser-based services and posting reports. The services SkyView at IRSA, SkyView at HEASARC, NCSA astronomical image archive, AstroBrowse, CDS's Aladdin, MAST, Vizier, and IRSA were all reviewed, at the discretion of FASST members.

ENABLING TECHNOLOGIES

The group membership was well prepared to discuss matters of scientific relevance and the needs of a scientist. However, the group did not take on the third question regarding the identification of technologies to be used throughout the archive services at STScI to allow our separate services to work together. This question is probably best answered by the software specialists in the archiving and catalog services. One relevant theme of the FASST recommendations is that the purveyors of archive software services (the various services that provide access to observation catalogs for all of the MAST missions, the Guide Star Catalog, and even the Sloan Digital Sky Survey) should enable a user to move seamlessly from service to service. Exactly what technology is required to enable this coordination was not discussed in FASST.

Technology that enables interoperability is changing quickly; MAST will have to rely on the expertise and background of its developers and advisors. One potential standard MAST might further investigate is Ed Shaya's XML-based XDF (extensible Data Format; http://xml.gsfc.nasa.gov/XDF/XDF_home.html). Archive developers should not work in isolation; they should be encouraged to learn about other projects at STScI and elsewhere, and to consider how interoperability between services and tools might be achieved.

SCIENTIFIC MOTIVATION

In this section, we provide scientific motivation for our study by generalizing some of the basic activities astronomers do to "do science". We describe three somewhat more specific scientific investigations. We then review the state of archive services in the community.

The classic scientific method involves testing, experimentation, and creation of further hypotheses. Each investigation provides increased knowledge or experience that potentially furthers scientific progress. The modern astronomer has opportunities and capabilities not available to her predecessors. She can collaborate with astronomers around the world with extraordinary ease now that email, data, and other files can be exchanged with ease over the Internet. She can access data taken from many different observatories from many different institutions such as STScI. She must frequently write and electronically submit proposals to support her research. She can quickly and easily access literature and preprints through existing services such as the ADS and astro-ph. She can access published catalogs through Vizier, which is supported by SIMBAD and ADC.

On the other hand, there are still difficult tasks and data challenges. Cross-correlation or cross-comparison of results obtained from queries of different services are not trivial, particularly in the case of very large result sets. In order to choose the most appropriate catalogs for her scientific requirements, in terms of data reliability, data quality, sky coverage, volume, frequency, and flux limitations, she must rely on experience and advice from colleagues in addition to literature searches.

Data discovery today often requires knowing that the catalog or the data exist before one sets out to find it.

Once the catalog is in hand, statistical analyses must be done with great care regarding the peculiarities and features of that particular catalog. "Outlier searches" done without understanding the data can be inefficient at best and misleading at worst. Given the ever-increasing volume of large public datasets, it is increasingly important to be able to assess the usefulness of such datasets without having to download a terabyte to a local disk drive. Documentation, description of units and conversions, and reliable error estimates are only a subset of what would constitute a catalog that is ready for constructing useful samples or for statistical analyses.

It also becomes increasingly important to be able to access scientifically useful and robust information from data without having to process many observations to a "science-ready" state. With HST Treasury GO programs and HST Legacy archive programs proposing to access and re-process and re-calibrate tens of thousands of WFPC2 datasets, MAST should be able to capture their results without forcing subsequent archive programs to repeat the preparation activities.

We first looked at a few scientific activities that FASST members would like to carry out and what is required to do them. This list of projects was intended to guide our discussion about the direction MAST might take to enable the best archival science.

1. Gather as much multi-wavelength photometry of galaxies as exists.
    a. Ask, given quality criteria, limiting magnitudes, sky coverage, spatial resolution, and overlapping bands, what exists?
    b. Create uniform catalogs (create source lists of reasonable photometry, filter bandpass, unit conversions)
        i. Receive a catalog generated from the original observations.
        ii. Receive data in a near science-ready format (accurate photometry and astrometry).
2. Starting with a catalog of quasars:
    a. Ask which of these quasars have observations available in the radio, optical, X-ray.
    b. Define the bandpasses of interest and the required data quality (signal to noise, for example) for the next query (c).
    c. Ask how many of these observations in found with (a) satisfy criteria defined in (b).
    d. Get and understand the observational data.
3. Derive the density in stars as a function of galactic location and stellar population age to look for changes in the localized star formation and its impact on the Galaxy
    a. Acquire multi-color (UV through NIR) stellar counts binned by galactic location.
    b. Get estimates of line-of-sight extinction towards these objects (from radio maps).
    c. Produce a catalog of these objects to do classification and age estimation.
    d. Successfully distinguish stars from galaxies.
    e. Where possible, gather multi-epoch astrometry to look for kinematic associations among stellar groups.
    f. Collect common photometry among different observatories/instruments.

These science cases are not intended to be the final goals of "FASST implementations", but to provide a sample of the goals of an NVO researcher, in the context of which MAST and other archives will operate over the next 5 years. We note a pattern of the scientific activities in the areas of data discovery, exploration, correlation, data filtering and data quality, and collection. These activities guided our recommendations and their prioritization.

We then set out to look at the state of services on the web that would enable such scientific activities. A catalog of information sites for astronomy was created (see the FASST website). Many committee members spent more than an hour trying to get information out of sources they had never used before, but would probably use for their science. From this exercise, we found that our ability to extract information from these sites was limited by the difficulty of (1) discovering the site and its services and (2) navigating through non-uniform interfaces. Mastering a single, moderate-size site, in many cases, took longer than the time we were willing to spend there. Further, finding specific details about the nature of the information (e.g. filter curves or instrument sensitivities) was very difficult if not impossible. Sites of note were MSX and 2MASS (both at http://irsa.ipac.caltech.edu), mostly for the excellent usability and the ability to get science-useable information without a lot of frills or mission-specific knowledge required. However, even with these excellent sites, members were left wanting to be able to search more than one catalog at once, to take the results of one query and make a query to another service, and to graphically analyze the information before or after filtering them.

While MAST has simplified the discovery of MAST data, these efforts only scratch the surface of the Institute-wide navigation and discovery problem. For example, one can now get information from the different mission collections in MAST, but accessing the response curves for the filters used to obtain the exposures is very difficult at best, especially for a non-local user. Not only is the filter information difficult to find, there is no standardized way to associate it with the archived data – thus there is no guarantee that it is available at all. Even once the metadata are understood, much effort is spent simply getting the results of a query into a new and distinct follow-up query. While the current MAST service that allows user-supplied catalogs for cross-correlation with the MAST catalogs has solved some of this problem, many other sites still require significant user intervention (if not retyping of data that are already in electronic form) to carry out such a simple task.

Most of the scientists in our group found they needed to spend significant time becoming an expert on several different sources of information and on learning to create scripts or other programs to make these sources work together. While making such activities easier and faster is also a motivation for the NVO effort, solving this problem locally should be a top priority. It is not our intent to suggest MAST create a service that supports the catalog needs of the entire astronomical community, but rather for it to develop a system that accomplishes interoperability for the archive query services and catalogs at the Institute and closely related sites. Coordination of catalog queries, if done right, could be extended to other archives that wish to join in the effort.

FASST RECOMMENDATIONS

In this second half of this document, we provide commentary on the following topics and we make our recommendations:

- The role of MAST in the NVO.
- The coordination of current MAST interfaces.
- Specific project recommendations.
- Implementation of software and data improvement efforts.
- Implementation of expanded data ingest and curatorial services.

THE ROLE OF MAST IN THE NATIONAL VIRTUAL OBSERVATORY

Here we make some general recommendations to MAST regarding the National Virtual Observatory.

As of the writing of this report, the National Virtual Observatory (NVO) is not yet a mature project, with a schedule and deliverables. However, there are some basic predictions we can make about the nature of the NVO that guides the philosophy of some of our recommendations. The NVO has two major manifestations: one that provides easy access to a broad variety of astronomical data and the other which provides distributed computing for advanced statistical studies of millions of data records. Arguably, the roots of the former manifestation are already in place, while the latter will require significantly more research and the presence of large, robust, and reliable catalogs. Enabling the former manifestation, we believe, will not impose a great burden on the existing NASA data centers. These data centers are now large and complex enough inside their confines that maintaining these centers in a cost-effective way requires self-describing, generic meta-data, user and programmer access to data search engines, and maps of their databases. To make such descriptions and access open to the astronomical community may be as simple as hosting the appropriate standard XML files that describe their meta-data and their query conventions and allowing http or Java servelet access to their query engines. MAST should expect to make some measured effort along these lines in coordinating such service descriptions with other archive centers.

However, MAST is expected to be a major node in the NVO. It houses some of the largest and most scientifically significant astronomical collections of data in the world. It is acknowledged by NASA to be the final repository of the optical and ultraviolet NASA missions, and it was chosen by the Sloan consortium to be the public release site for its data and catalogs. MAST serves the broad astronomical community as well as the optical/UV NASA community, providing and preserving specific expertise for both. MAST can tailor its services to serve the needs of both customer types, while recognizing that its unique appeal is in the arena of optical and UV astronomy, with a growing demand for near-infrared astronomy data and expertise.

MAST should, therefore, take the lead in NVO initiatives such as defining meta-data standards. MAST should initiate and carry out interoperability experiments and projects in collaboration with other data archives. MAST should be a conduit for scientifically significant contributions from individual astronomers, particularly those enabling further research and science using MAST datasets. MAST may find it most useful to take on specific scientific challenges and meet the scientific needs of these projects with services providing interoperability. Such a science driver would focus efforts. Institute scientists would provide a fertile resource for such scientific efforts, as will the up-coming HST Treasury and Legacy archive programs for HST's Cycle 11.

We apply this general recommendation to an immediate MAST question. MAST currently has two major interfaces, one is web-based and the other is Java-based. Both have advantages and disadvantages. *In order to be a major player in the NVO, MAST should not restrict development to only Java/Java-applets or to only web-based browser tools.* A full-service archive should be prepared to provide the optimal tool for the scientific activity. Java application or applet provides interactivity that would be nearly impossible in a browser alone; browser applications provide general access to the broadest community of users. Abandoning Java development would mean leaving graphical queries and other interaction-intensive features to other archives. Abandoning browser interfaces would leave some classes of users without access, and would discourage first-time users from visiting at all. We expand on this idea in the next section.

COORDINATING CURRENT ARCHIVE INTERFACES

For many years, there have been two main interfaces to the HST archive: StarView and the web interface. These two interfaces serve similar goals each with some differences and limitations. The Web interface to the archive is simpler to use, since it does not require software beyond a web browser. It is easy to find data that you know already exists, or to find data using coordinates and observation descriptions. To move much beyond such a search one must use StarView, a Java application, for which somewhat more comfort with a computer is needed to install. StarView creates a wider range of options that are not limited by the bounds of HTML forms, including customizable screen access to the full HST database. StarView can provide services that the web pages do not, in terms of interactivity and in the variety of query forms.

In the last year, the development of these two applications has become more coordinated, leading to improved functionality with less duplication of development effort. Coordination of future interface development should be guided by scientific objectives. Users prefer web interfaces for those activities that can be done in the context of a browser. External applications such as StarView or applets that work inside the browser may be required for functions requiring graphics and interactivity. For example, interacting with a FITS image in a web browser would require creating many new process requests, and would tax the network bandwidth to do things such as color-map stretching, panning, or zooming. Many applications should be developed in concert to share common services and goals, so one can accomplish the same ends from different starting points; either by using the same services or passing the user seamlessly from one program to the other. Finally, the coordination of applications should not be limited to StarView and the Web, as other applications such as SpecView and the tools of the APT may also serve to extend the functionality of the interfaces to the Archive.

SPECIFIC PROJECT RECOMMENDATIONS

In this section we provide a list of the recommended near-term projects and actions needed to prepare MAST to continue its leadership in the areas of space data archives and the NVO. Specific implementation suggestions and estimate efforts are discussed on pages 15-19. We refer to MAST here to include the Hubble Data Archive as well as the non-HST holdings and catalogs, such as the GSC2 and Sloan. These holdings and services are not under a uniform architecture, because each service is constructed and optimized for specific science goals. One of the main recommendations of the FASST group is to use this diverse architecture to model the realities of the NVO. In this document, when we refer to "data" we usually intend a broader sense of "data" than what astronomers usually mean. Data includes all information such as observations, catalogs, databases, or algorithms. Most of us have preconceived notions of the nature of these sources, notions that can limit the scope of the discussion. By expanding what we mean by "data", we allow for the more general cases of services that are not limited by our prejudices regarding the different sources of information. We also note that "services" are not always web pages, and we use the term "service" to refer to interfaces to the information, many of which can be accessed via the web, but which may also provide access to other applications.

The six projects with the highest scientific priority are listed in the Executive Summary. There are cross-references from that Table to this outline.

The following list of services or projects is organized into five categories, classified by the overall scientific goal for each of the services. Within each of the five categories, the projects were ranked within each category. Finally, the five categories were ranked. While this ranking is not a full ranking, it does reflect which goals are most important and identifies the five project categories that are the most important in moving towards that goal.

1) DATA DISCOVERY: Provide a centralized data discovery service. These projects included
   a) A "Yahoo" style portal service for STScI archive services and astronomical data (See http://www.excite.com or http://www.yahoo.com for examples of a general interest portal.). Since the STScI holdings and services are not currently housed under a uniform architecture; it reflects the reality of multiple services and holdings that exist external to STScI. The holdings of other data archives and catalog services do not follow the same conventions and standards. This portal will leverage the capabilities and services already provided by the MAST web site and StarView. An effective data portal will consist of at least four services:
      i) Data description: provide an easy-to-implement standard for a data service to describe its holdings, such as an XML file. At first, some cataloging of links and data could be done by hand but eventually, automated methods could be used once there was a large reference catalog or a number of XML descriptions of MAST services. There should be an AstroBrowse[1]-style method for getting all available information for a list of objects. There should also be a Google-style search capability (see http://www.google.com) for the contents of the portal as well. One potential method to determine the relevance of a

[1] AstroBrowse allows users to query hundreds of different astronomical catalogs and services around the world by sky position using a single form. The services can be limited by bandpass, data type, or keyword. AstroBrowse uses a metadata standard called "GLU", which is an ASCII file found in a configuration directory local to the AstroBrowse service. The next generation of AstroBrowse services may use an XML file, served from the individual services rather than a centralized configuration directory with a format that expands upon the standards set in the "GLU" files. (See http://heasarc.gsfc.nasa.gov/ab/ or http://archive.stsci.edu/starcast/ for AstroBrowse prototypes).

site to a specific search is to catalog sites based on the other sites they reference. Once such a portal is established for services here at STScI, such a portal could act as a model or even the foundation for a community-wide astronomical information portal.

ii) Query description: an easy-to-implement standard for a data service with catalogs of objects or observations to describe and stage its query services. Again, the solution could be in the form of an XML file describing the query service and how to access it. Each archive or catalog service would maintain its own query engine, which would be optimized for the types of scientific searches expected at each site. A description of the query service, its method of access (cgi call or a java servelet for example), query structure, and the mapping of some common query elements into their query structure would enable the most basic and common searches.

iii) Retrieval protocols. Data retrieval can involve multiple activities. An initial query of database contents could lead to a further query and retrieval of tabular data or the retrieval of associated data products. A retrieval result may also provide links to related data, tools, documentation, and software that allow the user to work with and understand the data.

iv) Expert domain definitions. STScI scientists and their associates have a wealth of domain expertise that spans a significant fraction of astronomy subjects. An expert can provide guidance on how to organize the data trees within his or her spheres of competence.

b) Lower priority and longer term (2003-2004): "my"-style portal customization capability.[2] Customization would allow users to set up accounts on the portal and to tailor the suite of services with the information they need to track all found in a single place. Examples of services a customized portal *might* provide (pending what scripts, Java servelets, and other services external archive centers offer at the time) would be:

i) Proposal and other deadline reminders.
ii) Proposal/data tracking.
iii) A quick search box for the MAST archive.
iv) Astro-ph or Astrophysics Data System (ADS) summary of recently submitted articles filtered by keywords supplied by the user, with links.
v) Lists of recent press releases from a configurable list of missions or projects.

c) Longer term (>2004): Agent based search/discovery tools. These tools would allow the user to supply a piece of software with some search criteria or information related to the status of a query or an observation and allow this agent to continue searching or monitoring and report back to the user when some criteria are met. The simpler of these agents may be web-based while ones that are more complex may require helper applications/plug-ins recommended in 2b below. The simplest example of this might be an agent that monitors the status of a data retrieval request and notifies the user, much like many mail tools tell the user they have mail. A more complex agent might monitor the archive and tell the user when datasets that fit some criteria become public so the user can request them.

2) INTEROPERABLE CATALOG SEARCHES AND DATA EXPLORATION: Coordinate catalog services. Provide the ability to use information from many unrelated sources at once. Provide the ability to explore the data via plotting tools and basic statistical analysis tools.

a) Coordinate the resources available under the MAST/STScI umbrella so that an archive researcher could tap into the GSC2 catalog, the Sloan catalog, and the MAST observation

---

[2] These "my" style portals allow different sites to be the masters of their own information, while the portal acts as a canvas on which information can be posted. An example of such services can be found at the OpenSource developers' information portal, Slashdot (http://www.slashdot.org), where the many different "slash-boxes" are administered by different sites but which are all accessible via the user customizable interface provided by the Slash software.

catalogs, and later expand to other archive services. This activity may involve proto-typing the helper application/plug-in recommended in 2b, but perhaps could be accomplished in another way.

b) Software helper applications/plug-ins to extend the functionality of the Web. The first tool of this nature could be a plug-in for a web browser that allows the user to easily take information from one web search and use that information as criteria for an unrelated search. One example might be to search Simbad (http://simbad.u-strasbg.fr) for objects of a particular nature and feed those coordinates into a MAST data search. A more complex search might require searching a catalog to find objects of a particular type, taking those coordinates to Simbad and getting the identification of those objects, then searching Simbad for *all* objects of that type, and finally going to the Sloan Digital Sky Survey and HST to get data for all those objects to analyze. This software would allow the user to mark the information from one query and to mark which columns of data are relevant. Then, once a search page is loaded in the browser the user could send the marked data into fields on this search form. Some astronomical knowledge, such as how to precess coordinates, might be needed to make this process work smoothly. Such a plug-in can also provide access or "hooks" to other applications, such as StarView or the APT, to extend the functionality of the archive interface, without forcing the user to re-start such extended activity outside of the browser. An external hook from a plug-in would provide a better interface to allow users to discover what they can (and may need to) do outside of the browser.

c) Allow queries that use multiple archives as sources of information for a single query. This feature might allow a search of 2MASS and GSC2 or SDSS object catalogs for objects with a particular infrared to visual flux ratio, defined by quantities already in those catalogs. The key would be to provide the metadata from each catalog that describes different characteristics such as beam size and pointing accuracy. This service would not attempt to create catalogs on-the-fly or "do science". If the catalogs already exist, however, a joint query could be allowed.

d) Create an easy-to-use visualization browser plug-in/helper application to take non-graphical information and plot it for a quick analysis. For example, one might want to plot a color magnitude diagram based on the text results of a search to find anomalous sources. The application could take the data displayed in text form in the browser window and use this plug-in to plot this data and even mark sources in the graphics window. If query results were returned in a standardized format (e.g. an XML file using a standard Document Type Definition (DTD) or even simple HTML tables), such a tool would be even easier to implement.

3) MAKE THE DATA EASY TO UNDERSTAND AND USE: Improve the archive's metadata to facilitate searches by people who do not have an expert knowledge of the missions in the archive or to provide information that is not available now.

a) Improve the coordinate solutions of existing HST images. The SHARE working group has also identified the improvement of coordinate solutions in HST data headers as one of their high priority recommendations. (Further discussion on page 16-17.)

b) Derive and save instrument-independent measures that quantify the data quality, such as a limiting magnitude or surface brightness. The methods for generating these numbers for HST data fall well within the realm of the SHARE working group, but we could establish standards for this information and help generate it for the other surveys and missions in the MAST archive. The MAST can take the lead in defining new quantities that may be necessary to describe observations or catalogs. For example, MAST could figure out a standardized definition of "limiting surface brightness" and its attendant modifiers such as (but not limited to) number of sigma, aperture size and shape, systematic error assumptions, and bandpass. Other simple quantities that are useful for archive searches probably require modifications to the calibration pipeline to compute image statistics and populate header keywords such as mode, number of cosmic rays, number of star-like objects, number of galaxy objects, and more. Pipeline modifications fall under the SHARE domain, but our committee acknowledged the usefulness of such quantities in finding suitable data in the archive, even if the quantities themselves are not directly usable for science.

c) Seek out and develop extensions to current meta-data standards such as the World Coordinate System (WCS). As an example, many have expressed the need for a nonlinear WCS but such a standard has not been implemented. There were recent FITS developments in October 2001 that now provide a roadmap for setting new WCS standards. MAST should support this effort by conforming to the standards and making suggestions, but significant support for the development of WCS standards by MAST is not expected.

4) CUSTOMER-RETURN AND CUSTOM USER SERVICES: Provide services that allow users to extend the functionality of archived data and provide facilities for people to add their own functionality and data.

a) Create a standard way of allowing users to submit information to be archived. Classic examples of a user-enhanced data collections and catalogs are the Hubble Deep Field survey and the Medium Deep Survey. Many people have used these data collections but their availability is more or less limited to users who already know they exist. The ability to ingest such collections of data into MAST and to allow users to discover and use such datasets or information derived from the datasets in the same way they currently can search for and print published papers at the ADS would be extremely useful. No formal refereeing process would be needed, aside from an assessment of the scientific usefulness of the data and its relevance to MAST. The data generally should be associated with a refereed paper or with a resource that is cited very frequently in the literature.

b) Make efforts to supply data from the archive in science-usable form. Currently most HST data are not clean enough to be science ready; the user must perform additional steps to get the data in that form. Leveraging the Institute's knowledge of the best ways to make the data science-ready would improve the usability of data for users who do not have such instrument specific expertise. For HST data, this recommendation falls within the realm of SHARE. Since MAST is an archive of other missions' data, such services are included in our recommendations. SHARE identifies such a service for co-adding HST images and spectra for a more "science-ready" product as one of their highest priorities.

c) Longer-term, also to be coordinated with SHARE recommendation: Allow users to extract data from images or spectra as part of the query and return only the extracted data. These extractions could be in the form of doing the photometry or in the form of sub-arraying the data to provide only the needed data rather than the entire array. For example, one could search for objects in all imaging data that falls within a particular bandpass. The search engine then extracts the photometry from each image based on the best methods of extracting such information from images from that instrument. The engine then returns only the photometry for analysis, rather than returning all of the images and leaving the extraction up to the user. The Space Telescope Users' Committee, we note, was wary of ventures too far along the path of data analysis. Object detection and photometry at the level required for top-notch science may be beyond the five year horizon unless the activity is integral to a peer-reviewed, fully funded scientific endeavor. Object detection and photometry at the level required for co-addition, however, could also be used to provide new keywords that would be useful in archive searches.

5) ACTIVE-TO-HERITAGE INSTRUMENT TRANSITION FOR HST: Create a standard policy and procedure for archiving the data and documentation for major missions and instruments, particularly for HST's instruments. This process should include input from the instrument teams. It is possible that the archive will prefer to make a custom policy for each instrument on a case-by-case basis; however, we recommend that a template be prepared in the near-term and then an agreement be reached between the archive and the instrument groups well

before the demise of the instrument and the group. This process should include flexibility for collaboration with the ECF and the CADC.

IMPLEMENTATION OF SOFTWARE EFFORTS

In the following paragraphs, we make suggestions regarding the technical implementation of some of the projects we recommend. We recognize, however, that the technical capabilities in this field are changing so rapidly that some of these specific implementations may be outdated by the time resources are available and allocated to these projects. While all of our recommendations are to be regarded as suggestions for integration into the MAST's strategic plan, we caution that this statement is especially true for specific implementation suggestions.

PORTAL AND SERVICE STANDARDS

The implementation of the portal projects reflects the implementation of the main portals on the web (e.g. Yahoo). We note that the order of their implementation is similar to how Yahoo came to be.

Starting with any number of portal software tools (either commercial or open source), we should look to construct such a tree-sorted portal for the information and facilities relevant to the MAST archive. With the diverse information sources found within MAST we should be able to come up with a good framework for a general astronomical portal. Even if a discovery portal were all this project achieves, deriving such a portal would be of great use to both internal and external users of MAST and perhaps the Institute as a whole. The creation and maintenance of this portal should take advantage of the local experts ("domain experts") who understand best how to organize the trees that fall within their realm. The development requirements are probably similar to what is invested in StarView now (about 2 FTE), with expertise in a broad array of web services and technologies. Domain experts may be able to contribute what is needed in about a week or two of work at 100% time. The final number of domain experts is difficult to assess. We would recommend using 1-2 experts to start, and then re-assess the project after their contributions are made. It is probably possible to create templates and interfaces from the first few experiences that will allow subsequent contributions by experts to be much easier.

Once this framework is in place, the next step would be to implement a Google-style search engine and cataloging scheme. Further, a broader implementation of AstroBrowse might satisfy the main search requirements for this portal. After this, we could see what sort of global community interest there is in such a site to see if it is worth pursuing funding to develop a field-wide portal in which user customization of a "my"-style site would also be included. Funding for broader efforts might come directly from NVO or tool development grants.

Creation of MAST data standards such as generic descriptions of catalogs, query protocols, and retrieval protocols will be a another step towards making MAST a full-service NVO data node. Such an effort will enable many archive interfaces to access MAST services. This work is already underway in MAST, and may require an additional year with 1 FTE level of effort to lead the way. Such efforts can start with organizing and describing the internal MAST services, but they should proceed in contact with other archives as the NVO conventions emerge and mature.

CATALOG COORDINATION AND VISUALIZATION PLUG-IN

Once the portal is completed, development of a complementary browser toolbar would be the next step. This toolbar could be similar to Yahoo's companion toolbar, which adds Yahoo-based functionality to Internet Explorer. This toolbar would install into a browser (most likely Netscape and other forms of the Mozilla browser that work under Windows, Solaris, and Linux). It would aid the user by filling out query forms derived from results from other queries and by making shortcuts to common features of the portal. For example, the user could highlight the name of an object on a web page, click a button, and see all the information about that object known by the portal or see what data in MAST exists for that object. The tool could handle precession of coordinates and even text reformatting (e.g. capitalization or changing from a comma separated list to a colon separated list). Implementation of visualization tools for plotting and basic statistical analysis could also be created for this plug-in, as well as tie-ins into other software developed at the institute, such as StarView, SpecView or the APT. Development of this tool is an additional 1.0 FTE beyond the 2.0 FTE for the portal.

DATA QUALITY, WORLD COORDINATE SYSTEM HEADER IMPROVEMENT

While these web projects are in development, we would strongly encourage the instrument teams to provide the archive with the instrument-independent data quality information for their systems, along the lines required to support the SHARE initiatives. By "instrument-independent information" we mean information that can be used by a Ph.D. astronomer who is not an expert HST user. Such information would be independent of the names of filters or gratings, for example. We could also pursue other experts for the other missions in MAST to help us derive similar numbers for those missions. Defining a quantity such as "limiting surface brightness" will also involve deciding on and communicating assumptions such as: number of sigma, aperture size, bandpass, and perhaps the technique or general equation used to create the derived quantity. This recommendation is thus in the form of recommending that the archive allocate resources for joint projects with other divisions, in the name of improving the quality and the "science-readiness" of the data. Such projects in the past have slipped between the cracks because no single division has the resources to make the projects a reality. A trial project on computing and providing limiting magnitudes or surface brightnesses for WFPC2 data, perhaps, could be accomplished with the cooperation of an archive scientist (10% time) with a WFPC2 scientist (10% time), a science data analyst (50-100% time), and the technical support (20% time) for the archive catalog in about two to three months.

We proposed several activities above, most of which overlap the SHARE group activities. Since we independently recommend these activities, we encourage MAST to support SHARE activities in leveraging the expanded processing to deliver improved and refined science products. We note that most of these activities require instrumental expertise for HST data that does not reside in MAST, such as the surface brightness project scoped in the previous paragraph. These projects, in general, require MAST input and coordination to implement, as well as interface work to proceed with awareness of the processing options.

The accuracy and the capability of the World Coordinate System keywords and contents for both images and spectra were of primary concern, since accurate coordinate information is vital to automated data stacking or data extraction. One of the most common requests for help with HST data analysis involves improving the astrometry of the images. The pointing solution for any given exposure is limited by the accuracy of the position of the primary guide star for that observation. Improved absolute solutions can be obtained by cross-correlating the positions of objects in the field with catalogs with very accurate positions such as Hipparcos, USNO, GSC2, or catalogs of radio standards. Improved relative solutions can be reached by cross-correlating objects in one frame with those in an overlapping image. Relative solutions are required for data co-addition and mosaics.

Implementing coordinate improvements in the headers of HST files may require a significant manual effort. The scientific goals of the improvements should be taken into account when planning the effort itself. It is also important to realize that the problem may require re-evaluations of the tradeoffs of scientific product vs. effort, if the effort turns out to be significantly more difficult.

- It is possible that bringing all of the data to a baseline absolute accuracy of ~0.5-1" is all that is needed to make data stacking more robust, which requires STScI to correct only the most egregious astrometric errors in the data.

- Accurate *relative* pointing solutions are the key for stacking images; relative pointing corrections are straightforward to obtain if the same guide star was used for the primary coordinates. However, if different guide stars were used, relative pointing solutions may require a first guess from the initial version of the header data, followed by relative position comparisons (cross-correlation) of sources common to at least two fields.

- Absolute astrometry to 0.5" or better may be possible for many HST images if comparison to the USNO A2 survey, the GSC2, or radio surveys are enabled. Such improvements would make catalog comparisons and image alignments of HST images much easier.

Revision of WCS standards and the improvement of HST data may require updating and adding to keywords of observation FITS files in the context of the On-The-Fly Reprocessing, with updates to the observation catalog as are done when other On-The-Fly Reprocessing improvements are added to the pipeline. Other requirements may include hardware augmentation. In order not to duplicate efforts, we will refer the reader to SHARE documentation for its recommendations and the level of effort required to improve the coordinate solutions for HST images.

INGEST SERVICES: SCOPE AND RESOURCE ESTIMATES

As a parallel development effort to the portal and its support tools, MAST should devise a standard method of submitting, cataloging, and providing permanent access to user supplied datasets and documentation. This activity includes defining standards and protocols for user data and for the transfer of inactive mission and instrument data. The planning for this activity should include the provision of easy, seamless, and system-wide access to this data. Such access perhaps could be modeled on the current MAST scrapbook effort but extended to cover a wider range of information.

INGESTING USER DATA

MAST should develop a process by which users can submit data for storage and distribution by MAST. It was generally agreed that user data could be very useful to other users, but quality control was a significant worry. The original data provider should be expected to provide the peer-reviewed journal articles or links to those articles. The data ideally should not be hidden away on some separate page or site. For example, if the data are originally HST data, those user-enhanced data should be easy to find by someone searching for HST data. In practice, this requirement means that user data ought to be listed alongside the HST data in the "science" table, or that a more global "science" table should be created that merges the information common to all MAST datasets. Additionally, it should be clear to users that the user data is not a pipeline product but is generated by an outside user. Thus, all of the warnings ("caveat emptor") apply. Literature links could be the primary method of documenting user data; however, MAST should investigate the means for also storing ancillary files such as README files. MAST should ensure that the data have sufficient documentation to be scientifically useful to an audience beyond the original contributors. MAST should not make ingesting user data any harder than it needs to be to maximize the scientific usefulness of the data. That is, it should not impose rigid extension structures, for example. Storing and distributing user data should be a straightforward, simple process, with most of the effort placed upon the user to provide the data, and the ancillary links and files needed for long-term viability of the data.

We recommend that MAST itself develop procedures and templates that will make acquiring and hosting user data painless, and the access of user data easy and scientifically productive. Metrics that gauge the success of a user data program could include: number of user programs, volume of user programs, development time required to support the acquisition of each user program (how many special cases required extra attention), the scientific productivity of the user program. We recommend that the "cost" of importing a dataset be very low, both to MAST and to the contributor.

The "user-data" effort will require a group of approximately five scientists and archive developers to meet over the course of about three months (perhaps requiring about 5% of their time for this period) and draft a policy for posting to the web. The coordination and review of such a document will require the attention of archive scientists, archive developers with database, web, and operational experience, and the instrument scientists of the instrument about to make the transition. Template and standards investigation should be done in the context of current MAST work, taking into account current standards and needs for templates. The group should review the current requirements for the replacement of the Data Archive and Distribution System (DADS) ingest process.

HERITAGE INSTRUMENT DATA AND SUPPORT INFORMATION

We recommend that each instrument team formulate a retirement plan, an "end game", well before the demise of their instrument. This retirement strategy should be developed in coordination with the archive facility, who might be expected to take on the final activities of, for example, running the data in batch mode through an On-The-Fly Reprocessing pipeline one final time or developing a long-term maintenance plan for reprocessing code. The complete retirement strategy should include not only a plan for dealing with calibration, but also a plan regarding the safekeeping and distribution of documentation, source code, binary code (and whether or not to maintain that code), and other ancillaries such as databases, instrument reports, and design requirements. The archive's goal beyond simply saving the data should be to make proper and informed interpretation of the data possible long after the individuals who created the documentation are no longer available. This coordinating work with MAST is in addition to the standard retirement and closeout activities of the instrument groups. It does not include any other closeout activities such as the completion of calibration programs. The coordinating work scoped here also does not include post-calibration efforts such as those undertaken by the ECF for some of the FOS data.

We recommend that a group be formed consisting of representatives from the archive and instrument teams to establish guidelines for instrument end-of-life activities and the transition to archival support for the dataset and its users. In addition, when a given instrument is nearing its end of life, a working group chosen from the archive and the instrument team should be established to implement this transition.

The group will address the following topics:

- What documentation should be turned over and in what form it should be? For instance, a synopsis document suitable for archival researchers should be created and be available along with selected instrument documentation and related information on the web.

- What software should be delivered and in what form it should be? The expectation is that some software sets (calibration, processing, analysis) should be made available as source code for documentation and reference.

- In what state the data should be? Beyond data formats and conformation to FITS standards, the data must be preserved for long-term access. A final reprocessing and calibration and a final repopulation of the metadata may be required eventually, since the processing software cannot be maintained indefinitely, and direct access to the calibrated data – particularly if it can be retained on-line – may speed up other extraction and mining services.

- What additional expertise can we preserve? We may wish to preserve contacts with instrument team members to help deal with questions from archival researchers; clearly such access cannot be maintained forever, but should be considered for some limited time.

The “instrument retirement” plans can be done over a long term, but should be accomplished at least 6 months before the demise of the instrument and the instrument team. The Faint Object Camera (FOC) may be suitable for the next opportunity. The production of an agreement that outlines the expectations of both the archive and the instrument team, including a list of products, a proposed schedule of work and allocation of resources should be modest, of order 5 people from AB, ESS, and HST (or their equivalents) meeting a few times. A lead would collect and integrate suggestions into a draft and finalizing a memo over the timescale of 3-6 months. Implementation will impact the work of at least two leads on each side of the transition. A small amount of extra work should be budgeted to coordinate with the ECF and CADC. We note that the investment in a prep effort will forestall and minimize the much larger, unplanned “emergency” work that will occur without such an agreement in place. We also note that this work estimate is in addition to the instrument team’s efforts to close-out an instrument, and that it does not include final calibration, further calibration investigations, or related activities.

CLOSING COMMENTS

The FASST team reviewed the current archive services inside and outside the Institute, and it recommended several near-term activities for MAST. Many of these activities can be integrated into the current MAST structure, as improving the procedures, guidelines, and the creation of templates and standards should make data management and distribution under MAST more efficient. The timescales include the time in which the service or policy would optimally be released for public consumption.

The portal and the catalog coordination recommendation includes software development, interface development, and coordination with other archives, perhaps through the Astrophysics Data Coordinating Council (ADCCC) or through collaborators on the NSF NVO project. MAST should consolidate its interface development so that the StarView developers, with their Java expertise, work closely with the archive scientists and the web developers in MAST, as well as the catalog experts from GSC2. Browser-based interfaces and Java interfaces can use the same web resources; coordinated development would proceed most effectively if all the developers were on the same team along with the scientists. The timescale for such services, however, should take into account the operational needs and priorities of MAST as well as parallel developments in standards at other archive centers. The first-priority portal and catalog services accessing STScI-only holdings and services, depending on the resources available, could be ready in less than a year. The browser catalog tool should be researched and scoped, before any serious commitment is made to its development. Tool development should proceed in coordination with similar efforts at other data centers, in order to leverage existing tools and to minimize duplication of effort.

The SHARE/FASST recommendations depend heavily on resources from the instrument groups to design and provide the extra processing software that extending the reprocessing power will require. MAST, however, will be expected to provide the interfaces through the browser and StarView to access those expanded processing capabilities. MAST will also be expected to facilitate any access to object catalogs that STScI has in-house. Therefore, the previous recommendation regarding coordination and consolidation of the access to STScI/MAST catalogs has high and early priority.

Finally, the FASST team recommends that MAST establish policies for the acquisition of user data and for the transfer of larger-scale mission data and instrument data collections and documentation. By writing down a policy in advance, MAST can allow users and instrument groups to make advance plans for the capture of user-enhanced data and to optimize the transition to heritage instrument support. MAST has already accomplished such transfers for missions such as IUE and HUT. That experience and the experience with the ad-hoc transfer of HST FOS data should inform this process. Furthermore, the SIRTF Legacy programs sponsored by the SIRTF Science Center at IPAC require them to have plans for hosting the results of SIRTF Legacy programs. MAST should investigate the SIRTF Legacy archival plans to see if anything is worth replicating. The FASST team recommends separating the policy creation activities, to distinguish the small-scale user contributions expected from Treasury and Legacy programs and the like from the large-scale instrument retirement plans that may require coordination with ECF and CADC and will certainly require extensive coordination with the instrument teams. The small-scale policy and associated templates should be available to users initially in the next 6 months, as Treasury programs begin to ramp up and plan their activities. The templates and standards can be improved as programs flow in and stretch the initial parameters over the next 2-3 years, but the initial policies and standards for a modest program for the return of user-enhanced data should be in place quickly.

PRACTICAL ACRONYM LIST

The following list contains most of the acronyms used in the document and an informal definition for each one.

AB Archive Branch, organizationally under ACDSD
ACDSD Archive Catalogs and Data Services Division
ADC astronomical data and catalog services
APT Astronomers' Proposal Tool, the new "Phase 2" proposal tool for HST
CADC Canadian Astronomical Data Center (host for HST data)
CDS, SIMBAD astronomical catalog services
DSS Digitized Sky Survey (digitization of an all-sky photographic survey)
ECF European Coordinating Facility, European counterpart to STScI
ESS Engineering and Software Services, a division at STScI
FITS a standard file format for astronomical data
FASST Future of Archive Services at Space Telescope (this report)
FTE Full Time Employee unit, here equivalent to percentage with a specified time frame.
GSC Guide Star Catalog
HEASARC X-ray astronomical data services

HTML Hyper text mark-up language (the markup language for web pages).
IRSA Infrared astronomical data and catalog services
MAST Multi-wavelength Archive at Space Telescope: UV/opt/near-IR astronomical data services
MOU Memo of Understanding
MSX and 2MASS : Major IR surveys hosted by IRSA
NCSA National Center for Supercomputing Applications, source of telnet and Mosaic, a browser.
NVO National Virtual Observatory
IRSA Infrared astronomical data and catalog services
HST Hubble Space Telescope
SDSS Sloan Digital Sky Survey
SIRTF Space Infrared Telescope Facility: will be renamed after launch
SHARE Study of the Hubble Archive and Reprocessing Enhancements (Gerry Kriss, chair)
STScI Space Telescope Science Institute
USNO US Naval Observatory star catalog
WCS World Coordinate System (a standard sky coordinate system, used in FITS)
XML eXtensible mark-up language (a markup language for data, informally).