A RECOMMENDATION BY THE HUBBLE SECOND DECADE COMMITTEE:

THE HUBBLE DATA ARCHIVE

Toward the Ultimate Union Archive of Astronomy

Space Telescope Science Institute
3700 San Martin Drive
Baltimore, Maryland
December 2000

INTRODUCTION

The archive of data from Hubble is undoubtedly the largest and most heavily used collection of pointed observations in astronomy today. Its volume, presently over 7 terabytes, is increasing at a rate of over 100 gigabytes per month. Meanwhile, the rate at which scientists withdraw data from the archive has swelled to about 15 gigabytes per day, which is four times the ingest rate of new data from the telescope.

The archive is moving to the center of the Hubble program. Its increasing volume and usage—as well as the new approaches to astronomy it empowers—assure its growing importance in Hubble's second decade. In 2010, after the end of telescope operations, Hubble research will become totally archival. At that time, the Hubble harvest will be viewed as several 'crops'—published results, educated students, an informed public, and a wonderful archive of data and new modes of archival research.

The new modes of research involve vast amounts of information, often located at different sites, being queried, correlated, downloaded over the Internet, and analyzed on desktop computers by astronomers around the world. Computer and communications technologies are advancing. The quality and consistency of astronomical data are increasing. Falling are the artificial barriers that once partitioned astronomy according to techniques and wavelength regimes. Emerging are wholly new research opportunities based on the ability to compare and combine large sets of previously unrelated astronomical observations. Hubble is a leader for the community in these developments.

The Second Decade Committee recognizes and supports the core responsibility of the archive to capture Hubble data and deliver it to users. Even those tasks demand ongoing upgrades as the archive grows and technologies become obsolete. It is possible that

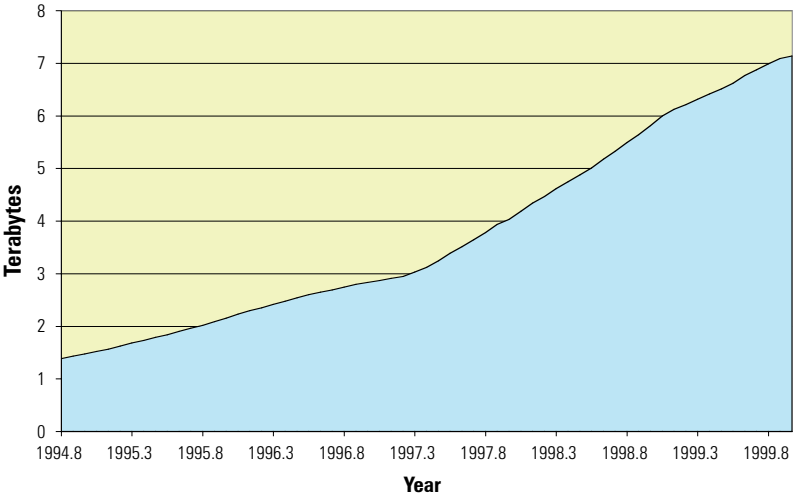

Figure 1. The cumulative data holdings of the Hubble archive as a function of time over the last 5 years. The increase in growth in early 1997 corresponds to the deployment of STIS and NICMOS. The small decline in growth rate in late1998 corresponds to cryogen depletion in NICMOS. The rate of data ingest will increase sharply with the installation and subsequent parallel operations of ACS and WFC3 in 2001 and 2003.

resource requirements to hasten the day of visionary archival research will compete with allocations to make or improve new Hubble observations. In that case, the Committee would favor new or better data over accelerated archive enhancements. However, we believe the potential for such programmatic conflict is minimal and manageable. Indeed, highest-quality data is top priority for both Archival Researchers (ARs) and General Observers (GOs).

We support the burgeoning phenomenon of archival research and the Hubble archive's leadership role in the new field. Indeed, we view the concept, design, and implementation of the Hubble archive as a significant intellectual contribution in itself. Our primary recommendation—which will be rendered more understand-

able by the review of the archive's special history and mature status, below—is to proceed apace with promising developments underway. Our caution is to remain prudent about which developments to pursue with Hubble resources and which to await from Hubble's partners near and far, who build with us toward the ultimate, union archive of astronomy.

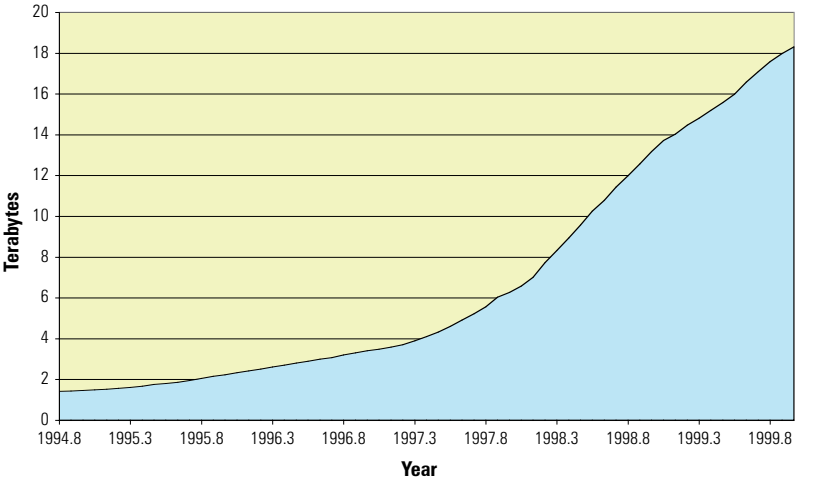

Figure 2. The cumulative data volume retrieved by astronomers from the Hubble archive as a function of time. As of November 1999, just over 18 terabytes had been extracted, or just over 2.5 times the data volume ingested up to that time.

HISTORICAL ROOTS OF AN EMERGENT VISION

In the late 1970's, the Space Science Board (SSB) of the National Academy of Sciences developed concerns about the preparedness of the NASA science community to accommodate the flood of new data expected from space missions under development, including the Hubble Space Telescope. With computing capabilities advancing rapidly, and space scientists becoming more sophisticated com-

puter users, the SSB felt a comprehensive strategy and guiding vision were needed to match sources and users of data in the most effective and efficient manner. To this end, the SSB chartered the Committee on Data Management and Computation (CODMAC) to study the issues and formulate recommendations.

The most fundamental and far-reaching recommendation of CODMAC was that scientists should be involved in all stages of data management for space science missions, from acquisition to archiving to distribution. CODMAC felt that an organization with a vested scientific interest and hands-on expertise in the data would best assure the integrity and usefulness of the data. In retrospect, this pivotal decision led directly to the development of the distributed data management architecture that exists today in astrophysics, planetary science, and space physics.

Hubble presented an early opportunity to apply CODMAC's strategic vision. When work on the Hubble archive began in 1984, the Institute was responsible for it, rather than the National Space Science Data Center (NSSDC).

Hubble is an international project of NASA and the European Space Agency (ESA), and involves scientists and the scientific institutions of many countries worldwide. This international character of the Hubble science program is epitomized in the geographical distribution and cooperative development of the archive.

The terms of the original Memorandum of Understanding (MOU) between NASA and ESA required that a full copy of the Hubble archive be established at the Space Telescope European Coordinating Facility (ST-ECF) at the European Southern Observatory (ESO) facility near Munich, Germany. The purpose of the ST-ECF archive copy is to support data distribution to European astronomers. In Canada, limited international network connectivity led to establishing the Canadian Astronomy Data Centre (CADC), which hosts Hubble and other archival data sets of

interest to Canadian scientists. In the past year, a fourth Hubble archive site (for non-proprietary data only) has been established at the National Astronomical Observatory of Japan (NAOJ). Procedures are now in place for the Institute to provide ST-ECF, CADC, and NAOJ with copies of Hubble science data on a routine basis.

ST-ECF and CADC participated with the Institute in the design of the prototype for the Hubble archive, which was called the Data Management Facility (DMF), and was later superseded by the Data Archive and Distribution System (DADS). For more than 15 years, the Institute, ST-ECF, and CADC have held regular coordination meetings to share experiences and set goals for the archive. With the Institute shouldering the bulk of the day-to-day operational responsibilities, ST-ECF and CADC have explored alternative, innovative data access and delivery mechanisms. Efforts to upgrade storage media are a first example of technical cooperation among the Hubble data nodes. DADS stored Hubble data on 12-inch 'write once read many time' (WORM) optical disks. ST-ECF and CADC migrated to Compact Disk Read-Only Memory (CDROM) data storage. At the current time, ST-ECF is evaluating Digital Video Disc (DVD) technology as an archival medium. Meanwhile, the Institute plans to migrate to magneto-optical (MO) storage, which is a mature yet growing technology with a large installed base.

Astronomy Archives & Information Services Online

ADC	*http://adc.gsfc.nasa.gov/adf/*
ADF	*http://hypatia.gsfc.nasa.gov/adf/*
ADS	*http://adswww.harvard.edu/*
CADC	*http://cadcwww.dao.nrc.ca/*
CDS	*http://cdsweb.u-strasbg.fr/*
Chandra	*http://asc.harvard.edu/cda/*
ESO	*http://archive.eso.org/*
HEASARC	*http://heasarc.gsfc.nasa.gov/*
Hubble	*http://archive.stsci.edu/*
ISO	*http://isowww.estec.esa.nl/*
IRSA	*http://irsa.ipac.caltech.edu/*
MAST	*http://archive.stsci.edu/mast.html*
NED	*http://nedwww.ipac.caltech.edu/*
NAOJ	*http://www.nao.ac.jp/*
NSSDC	*http://nssdc.gsfc.nasa.gov/astro/*
ST-ECF	*http://ecf.hq.eso.org/*

MO technology is comparable in cost to current-generation DVD, has proven long-term stability, and has better speed performance than other optical media. (The Institute had planned to migrate to DVD as a storage medium in the expectation that DVD would quickly supersede CDROM. However, at the time we needed to make a selection of new media, the industry has yet to settle on a standard, and the Institute was concerned that selection of one DVD format over another was too risky.) NAOJ is using CDROMs for data storage, and the Institute is using its bulk CDROM production system to back-populate that archive.

The Institute, ST-ECF,and CADC have pursued collaborative and complementary efforts on the Hubble archive in several other areas as well:

CADC developed facilities for automatically generating preview images for data (after the proprietary period has elapsed). These previews are now made available to all of the archive sites. The SkyCat data visualization tool, which includes automated network access to distributed catalogs, was developed by ESO/ST-ECF and CADC together.

ST-ECF provided the first full-function interface to the archive, called STARCAT. The Institute developed the X-Windows-based StarView interface, the first major Hubble software package to be developed using object-oriented methodologies and compilers. StarView could be used both locally and distributed to users for installation at their respective institutions. ST-ECF and CADC developed World Wide Web (WWW) based interfaces to the archive and complementary catalogs (e.g., WDB, the web-to-database query interface developed at ST-ECF). Web browsers, which came along several years after StarView, still provide a sufficient interface to the archive for many users. The Institute has recently developed StarView II, a successor to StarView. Like StarView, Starview II provides sophisticated query screens of StarView, which

are not possible with web-based forms. These query screens enable new levels of interactivity with the archive and the associated catalogs. ST-ECF contributed to StarView II by providing Java preview display modules.

The Institute undertook the Hubble Archive Re-engineering Project (HARP) to reduce archive operations expenses and to extend the useful lifetime of the DADS optical disk-based archive system. Key elements of HARP were data segregation (moving engineering and other less frequently used data to separate media), data compression, and data migration to lower-cost, higher-capacity, magneto-optical media.

CADC and ST-ECF took the lead in developing on-the-fly calibration (OTFC) to economize on storage costs and make the highest quality data available to archival researchers. Uncalibrated data compresses much more efficiently than calibrated data, and OTFC means that only uncalibrated data need be archived. For the scientist, OTFC means that data retrieved from the archive has been processed using the best possible calibration reference data. Thus, OTFC minimizes the possibilities of misinterpretation of the data and removes the burden of recalibrating data manually. Based on the popularity and lower operating cost of the CADC/ST-ECF OTFC system, the Institute has implemented an OTFC facility enhanced to accommodate changes in data formats and calibration algorithms.

ST-ECF extended the concept of 'data associations' to WFPC2 observations. Data associations are data sets that must be processed together, a concept originally developed for STIS and NICMOS data. The extension to WFPC2 now permits the automated alignment, cosmic-ray cleaning, and co-addition of multiple images of the same field.

The Institute, CADC, and ST-ECF all provide access to other space—and ground—based data archives with a large integrated

capacity. CADC also hosts data from the Canada-France-Hawaii Telescope, the James Clerk Maxwell Telescope, a copy of the Institute's Digitized Sky Survey (DSS), and provides access points to a number of other astronomical archives. ST-ECF provides coupling to ESO's archive, which includes data from the New Technology Telescope, the Very Large Telescope, and the Wide Field Imager on the ESO/MPIA 2.2-m telescope in La Silla. The ESO archive will include data from the VLT Survey Telescope once it becomes operational. The Institute operates the Multi-mission Archive at Space Telescope (MAST), which is NASA's UV/optical/near-IR archive center. MAST includes data from the International Ultraviolet Explorer (IUE), Astro's Hopkins Ultraviolet Telescope (HUT), Astro's Ultraviolet Imaging Telescope (UIT), Astro's Wisconsin Ultraviolet Photopolarimeter Experiment (WUPPE), Orbiting Retrievable Far and Extreme Ultraviolet Spectrometers' (ORFEUS's) Interstellar Medium Absorption Profile Spectrograph (IMAPS), ORFEUS's Berkeley Extreme and Far-UV Spectrometer (BEFS), and Copernicus. MAST provides direct access to the data of the Extreme Ultraviolet Explorer (EUVE), and will include data from the Far Ultraviolet Spectroscopic Explorer (FUSE). MAST supports the DSS, the Very Large Array (VLA) Faint Images of the Radio Sky at Twenty centimeters (FIRST) survey, and will support the Mosaic Imager of the National Optical Astronomical Observatories (NOAO). Through the Astrophysics Data Centers Coordinating Council (ADCCC), the Institute works closely with other archives and services to increase interoperability and to provide increasingly transparent access to distributed astronomical data holdings. Other participants in ADCCC include the NASA-sponsored High Energy Archive Science Research Center (HEASARC) at Goddard Space Flight Center (GSFC), the Infrared Science Archive (IRSA) at Caltech's Infrared Processing and Analysis Center (IPAC), the Advanced Xray Astronomy Facility (AXAF, now

Chandra) Science Center at the Center for Astrophysics (CfA), the NSSDC (GSFC), the ADC and Astrophysics Data Facility (ADF) at GSFC, and the Astrophysics Data System (ADS) at CfA. The Institute also participates in NASA's Space Science Data System, which aims at interoperability across all space science disciplines.

The Institute, CADC, and ESO/ST-ECF all have close ties with the catalog and bibliographic services provided by the Centre de Données Astronomiques de Strasbourg (CDS) and NASA Extragalactic Database (NED). The Institute and HEASARC have led the development of AstroBrowse, a cross-archive data search and discovery tool based on CDS's AstroGLU system for locating astronomical information. The ADCCC has partnered with planetary science and space physics data providers to develop a successor to AstroBrowse, called Interoperable Systems for Archival Information Access (ISAIA). Not only will ISAIA locate data of potential interest to the user; it will integrate the query results from multiple sites and services into a single format.

Today, the stable, mature facilities of the Hubble archive provide an arena for two exciting kinds of exploration. First, the archive is used intensively for scientific research and discovery. Second, the archive is in the vanguard of efforts worldwide to improve the art and practice of archival research itself. This dual role—serving science while advancing the means of science—resonates with NASA's original decision to involve the astronomical community at the heart of the Hubble enterprise. This decision is paying off now in the archive just as it has in two other conspicuous areas—Hubble's ever more efficient science operations and ever improving on-orbit instrumentation.

INTEROPERABILITY OF ASTRONOMY ARCHIVES

Increasing interoperability expands the data grasp of scientists by permitting observations of the same objects from various missions and instruments to be gathered and correlated. This expanded spatial, spectral, and time coverage provides a more synoptic view. As a natural extension, links between data sets and the published literature can inform the researcher of previous studies of the same objects or work with the same data.

MAST already enables multi-wavelength, multi-mission correlative science. Using WWW interfaces, the user can search for data on a given astronomical source from various instruments and missions, and, for multi-wavelength data, for classes of objects in astronomical catalogs. Further developments now underway will provide cross-correlations between the catalog of Hubble observations and other object catalogs via the ADC's generic interface to its catalog collection.

The breakthrough power of interoperability is best illustrated in the case of surveys.

Historically, astronomers have relied on optical spectroscopy to classify definitively the sources discovered by sky surveys. Before the advent of multiplexed spectrographs, this task was so laborious as to be infeasible for surveys producing more than a thousand or so sources. For example, the Extended Medium Sensitivity Survey (EMSS) by the Einstein Observatory found

Toward the Ultimate Archive

• Strengthening connections to other archives, catalogues, and abstract services, for broader research parameter space and links to the literature.

• Advancing technologies for computers, networks, data compression, and storage media, to retrieve and analyze more information more readily and at lower cost.

• Improving calibrations and creating more higher-level data products, to make data more science-ready.

• Data mining with new software tools and new catalogs of object properties, to enable higher-order research based on questions posed in scientific terms.

835 serendipitous X-ray sources over 780 deg2, complete down to a flux ~10^{-13} erg cm^{-2} s^{-1} in the 0.3-3.5 keV band. Because of the large positional error boxes, each X-ray source had typically several candidate optical counterparts, of which spectra had to be taken one at a time until a reasonable identification and classification could be made. The EMSS took about 10 years to complete (Stocke et al. 1991, *ApJS* 76, 813).

With dedicated telescopes, multiplexed spectrographs, and automated procedures, classification by optical spectroscopy can be extended to far larger surveys, but only when the sources are sufficiently dense on the sky and sufficiently bright. These factors apply to the Two-degree Field (2dF) survey, which will classify about 250,000 sources, and for the Sloan Digital Sky Survey (SDSS), which will classify about 1,000,000. Both surveys will be limited to about 20th mag.

Spectral classification is infeasible for large surveys when one or more of the enabling factors—abundant funds, dedicated facilities, high density of bright sources—is missing, as is usually the case. Four examples: the ROSAT All Sky Survey (RASSBSC) includes 18,811 sources over 92% of the sky. The White, Giommi, and Angelini (WGA) catalog of ROSAT Point Sources includes about 70,000 sources over ~10% of the sky. The NRAO VLA Sky Survey (NVSS) includes almost 2 million radio sources north of a declination of -40º. The Faint Images of the Radio Sky at Twenty cm (FIRST), which probes a factor of 3 deeper than NVSS, will contain ~900,000 sources over 10,000 deg^2. For these surveys—even restricting their scope to sources brighter than ~20th mag—the resources do not exist to perform spectroscopic classification in a reasonable timeframe.

As sources become fainter, classification based on multi-wavelength, statistical methods becomes the only option. A 4-m-class telescope can identify a source with relatively strong features (such as

a quasar) in a 1-hour exposure down to 22nd-23rd mag; a 10-m-class telescope can reach 25th-26th mag in about the same time. However, these limiting magnitudes are inadequate to identify, for example, the HDF's faintest objects spectroscopically. Deep surveys at other wavelengths encounter the same problem. A typical XMM exposure, for example, will reach X-ray fluxes fx ~ 10^{-15} erg cm^{-2} s^{-1}. At those levels, calculations using the appropriate X-ray-to-optical flux ratios show that all radio-loud AGN are fainter than the 4-m limit for spectroscopic identification. Most will lie beyond the 10-m limit. As a second example, even most radio-quiet AGN found at the Chandra limit (fx ~ 10^{-16} erg cm^{-2} s^{-1}) will be so faint as to require exceedingly long integration times for optical identification using a 10-m class telescope. And normal galaxies found at the Chandra limit, which will be relatively brighter in the visible, will remain problematic for spectral classification due to their lack of strong spectral features. A final example is provided by the FIRST survey in the radio band. Most radio-loud sources found at FIRST's 1 mJy limit will be ~24th mag in the visible, which means that a 4-m telescope will not be capable of identifying them spectroscopically.

Such massive, deep surveys demand an alternative method of classification, and interoperating, multi-wavelength archives can provide solutions. For rare populations, an approach that works well is cross-correlating catalogs in different wavelength bands to preselect candidates based on spectral energy distribution. Optical identification can then proceed on a smaller, tractable sample. One example of this approach is the 'photometric redshift' selection of the most distant galaxies in the Hubble Deep Field, which searched for UV 'dropouts' in WFPC2 images obtained through four filters. Another example is the Deep X-ray/Radio Blazar Survey (DXRBS) (Perlman, Padovani et al. 1998, *Astron. J.* 115, 1253), which cross-correlated the WGA catalog with the GB6 and PMN catalogs (~120,000 sources), finding ~1,600 objects that were both X-ray

and radio sources. A further down-selection based on radio spectral index reduced the sample to ~300 objects, ~95% of which turned out to be blazars.

The Second Decade Committee supports establishing closer ties with other archive centers to fully exploit the multi-wavelength as well as temporal parameter space offered by those collections. We support the development of catalogues of Hubble sources based on homogeneous subsets of the HST archive and with clearly defined quality metrics. These can be especially useful in conjunction with the products of other astronomical catalog sites, and provide a baseline model for the archival services for NGST and other future missions.

We support closer links with abstract services, such as the one provided by the Astrophysics Data System, to connect data and astronomical papers.

COMMUNICATIONS & COMPUTER TECHNOLOGIES

Advancing technologies are speeding access to data, and enabling larger volumes of information to be acquired and managed. Feeling the pressure of larger data sets and the lure of new opportunities for research using interconnected archives, HST users will continue to demand access to these technical advances. The challenge for both users and archive operators is to purchase new technology when the price is right and the choice is clear.

For the end users, the most visible archive technologies are involved with delivering HST data to their desktop computers. Currently, that delivery is via Exabyte tape or the Internet; both methods are becoming inadequate. To illustrate, consider a set of observations using the Advanced Camera for Surveys (ACS) after 2003. The typical data volume of a GO ACS program will be 3 to 9 gigabytes, larger by almost a factor of three than a typical WFPC2 observing project's data volume today. Fortunately, data-storage

technology has kept pace with the growth of the archive and data-delivery demands. Indeed, a philosophy of 'planned obsolescence' on 5- to 7-year timescales assures the long-term viability of the archive. Thus, DVD and data compression, which are near-term developments, will be superseded by newer technology by 2010.

The Second Decade Committee encourages targeted research and sagacious investments in data-delivery technologies. We favor improvements in network delivery over hard media because of the potential cost savings and the overall advantages of the WWW as a nexus of archival research. The current inbound bandwidth to most researchers' computers (~100 kilobytes/sec) would not be adequate for the timely Internet transfer of a typical ACS observing set. In a few years, however, inbound bandwidths of 100 megabytes/sec should be the norm, which is adequate for such high-volume transfers. (The Institute outbound bandwidth today is ~20 megabytes/sec, and upgrades to 100 megabytes/sec are planned.) The Committee further recommends that lossy data compression be explored, which could reduce data volumes by perhaps a factor of ten with negligible loss of information in many scientific applications.

More science-ready data

Hubble introduced the era of vast amounts of consistently high-quality, multi-faceted astronomy data. Today, the data in the Hubble archive—as well as that in the archives of other space missions and concerted ground-based survey programs—is already more science-friendly than data of earlier times. This development is due to more precise instruments, more constant observing conditions, and more consistent data processing. The Hubble archive will become even more powerful as its content is made more science-

ready by improving calibrations and adding more higher-order data products.

Today, Hubble operates with its second generation of instruments, and a third generation will be installed in the second decade, including ACS, COS, and WFC3. In time, the calibrations of earlier-generation instruments will cease to improve, aside from possible changes in fundamental reference data. At some point, the long-term cost of maintaining calibration software will exceed the cost of archiving a final calibrated data product. Anticipating this crossover point, the ST-ECF staff have been working on strategies for final calibrations of FOS and GHRS data. Similar plans must be made for the other instruments.

The best possible calibrations are necessary, but they are not the only data products astronomers need for the most powerful analysis of Hubble data. Data characterizations and catalogs of selected data sets are extremely useful, as the HDFs and Key Projects have demonstrated. As more large observing projects and surveys are undertaken, the value of such investments in accessibility and understandability will increase further.

As the Hubble archive improves in both data quality and access, there will emerge new opportunities for 'virtual surveys' to discover rare, relatively bright objects. Whereas deep, pointed observations like the HDF are useful in detecting faint but rather common objects, they are not useful for the random discovery of rarities, such as distant supernovae, blazars, or clusters of galaxies. Virtual surveys will sift vast archive holdings in search of such rare objects. Prominent among such holdings will be the avalanche of deep images from the ACS operating at high duty cycle in parallel mode. The exposures from deep, wide-area, ground-based, CCD mosaic surveys comprise another such trove. The potential of archive-based searches is illustrated by the discovery of several $z = 0.4$ clusters of

galaxies in the (parallel) observations of the Medium Deep Survey (Ostrander et al. 1998, *Astron. J.* 116, 2644).

Another powerful advantage of the Hubble archive (and the archives of other space missions) is access to an enormous amount of data acquired uniformly under extremely reliable conditions. For example, a morphological study of a statistically complete sample of 341 galaxies with measured redshifts in the range 0.3 - 0.9 (Brinchmann et al. 1998, *ApJ* 499, 112) was based on image data obtained from three independent Hubble programs, including the Groth strip (Groth et al. 1994, *BAAS* 185, 5309). These data could be calibrated and intermixed consistently, which is essential for building large uniform samples. The systematics of various methods of classifying galaxies could be tested and quantified, a feature lacking in studies where the data and the methods are unavailable outside the author's domain.

The Second Decade Committee supports the improving calibration of Hubble data, culminating in a definitive calibration. We encourage the development of higher-level data products for the archive, especially to support research on the results of large, uniform observation sets, such as produced by the Hubble Treasury Program and routine parallel observations. Even for smaller observing projects, we see a benefit in encouraging researchers to provide to the archive such data products as might help future researchers working with the same data. (The GO's final data products will sometimes have more value to the archival researcher than the basic calibrated data currently provided.) The Committee also sees a value in the Hubble archives serving as repositories for large processed data sets described in published scientific literature.

DATA MINING

The forces shaping the Hubble archive of 2010 are now visible. They are increasing interoperability, advancing technologies, and improving calibrations. These three force vectors point toward a new kind of research, which currently has the label 'data mining'. Today, data mining is more a vision than a clear picture, but it is compelling. The idea is that users can pose high-order questions to a distributed information system comprising multiple sources of data. This grail is the ultimate archive, and this concept has now been endorsed as the Virtual Observatory in the National Academy of Science Astronomy and Astrophysics Survey Committee's report on priorities for the next decade.

Two types of development are needed to hasten the era of data mining. First, data characterizations and data associations must be developed as a 'switching yard' between the user and the data. In the case of Hubble, this would involve characterizing the objects in images and spectra, and creating a database of such characterizations. An example involving imaging data would be the object characterizations resulting from analyzing the Hubble Deep Fields (HDFs) or the Medium Deep Survey. For spectral data, an example of an association might comprise all spectra for a given object, grouped as a single data set, with metadata to describe the spectral resolution, wavelength coverage, and signal-to-noise ratio. Developing a pipeline that extracts meaningful and useful object attributes from the highly heterogeneous collection of Hubble data will be a substantial challenge, but the potential benefits of such a facility are enormous.

The second required development is software tools to formulate queries and access data through the characterizations and associations located at different sites. These tools are simply concepts

today, but the hope is bright and renders the future of archival research lustrous indeed. One can envision users posing directly, in scientific terms, such queries as, 'Are there clusters of galaxies with X-ray fluxes of at least 10^{-13} erg/s/cm2 at the positions of these steep spectrum radio sources that were also observed by imaging cameras on board HST?'

The ST-ECF, in conjunction with ESO and the CADC, is conducting a data-mining pilot project based on associations for images from Hubble's WFPC2. The database of associations will contain an object list (positions, magnitudes, object shape parameters), statistics on the object list itself (number of each type of object, magnitude distributions, etc.), the limiting magnitude for the association, background characteristics, lists of objects in the field of view from GSC I and II and from other Hubble observations, and associated Point Spread Functions (PSFs).

The Second Decade Committee supports developments leading to data mining, but feels that Hubble resources should be used for cooperation rather than leadership, which should be ceded to outside initiatives. We support the emergence of long-term goals for the Hubble archive, which should be developed in common by the many organizations supporting the distributed archive. A cooperative plan should draw upon the strengths of each organization and also other major players in scientific archiving and computer science, where various forms of 'data mining' are subjects of both pure and applied research.

COMMITTEE MEMBERS

Stefi Baum

James Beletic

Robert Brown, chair

Tim de Zeeuw

Larry Esposito

Michael Fall

Robert Fosbury

Richard Green

Timothy Heckman

Garth Illingworth

Shrinivas Kulkarni

Henny Lamers

Mario Mateo

John Mather

Claire Max

Donald McCarthy

Richard McCray

Keith Noll

Ethan Schreier

Charles Steidel

John Stocke