The Virtual Solar Observatory

A White Paper

25 May 2000

Further information available on request to fhill@noao.edu

Executive Summary

The Virtual Solar Observatory (VSO) is a new tool for investigating the physics of the Sun and its impact on the Earth environment. The VSO will address one of the central challenges of solar research: the need to locate, correlate, absorb, and analyze data from a wide array of scientific instruments that measure the Sun on spatial and temporal scales that range over seven orders of magnitude. Currently, this process is extraordinarily time-consuming and labor-intensive; yet it is pursued, because the coupled Sun-Earth system demands it. The VSO will greatly increase the power and pace of the correlative studies needed to address fundamental challenges such as predicting geomagnetic storms, understanding solar irradiance variations and their effect on the Earth environment, detecting active regions below the solar surface before they cause solar activity, and understanding the generation of the solar wind

The VSO is a scalable environment for searching, integrating, and analyzing databases distributed over the Internet. Initially, the VSO would comprise a federated system of 10 solar archives containing data from space- and ground-based instruments, from US and international sources, from NSF- and NASA-funded groups. A key element of the VSO is an integrated data-mining and analysis capability that can be applied both across and within databases.

The VSO will make superior use existing databases that represent a large investment of time, money, and expertise. The capabilities of the VSO will be available to a wider community of scientists, forecasters, and the public. The VSO will be a natural conduit for education and public outreach in solar and space physics.

The management of the VSO will be via a small core team and a Scientific and Technical Oversight Board. This structure provides accountability as well as distributed expert oversight.

It is estimated that the VSO could be placed into operation within two years at a cost of $7.2M and then operated for $1.7M per year, based on 10 component archives.

1. Introduction

The Sun exhibits complex phenomena, such as flares, that pose important challenges to our understanding of the universe and which can directly influence our increasingly technological civilization. Solar structure varies on time scales from microseconds to decades and on horizontal and vertical spatial scales ranging from 10 to 10^6 km. Understanding solar structure, and ultimately predicting solar behavior, requires the correlation of many types of observations. These observations, which cover the entire electromagnetic spectrum as well as energetic particles, cannot be obtained with any single instrument or indeed at any single observatory. Thus, the data are stored in a myriad of locations, in dozens of formats, and analyzed by numerous applications. A researcher who wishes to combine several data sets faces the tasks of finding the data somewhere in the world, searching each set for the observations that most nearly fulfill the precise scientific requirements, arranging to have the data delivered, and mastering the various storage formats and calibrations. These labor-intensive steps must be completed before any analysis can even begin, and have deterred many researchers from performing studies that would provide valuable new insights into solar structure and the mechanisms of solar activity and variability.

In spite of the difficulty of using current data systems, solar physics has long exploited the scientific richness of cross-correlative studies. This is evident from a brief survey of the articles published in volumes 184 and 185 of *Solar Physics* which contain 39 papers with observational results. Of these 39 papers, 27 used from 2 to 33 disparate data sets. On average, 4.5 data sets were used in each of the 39 papers. While this is a very small sample, it is indicative of the importance of multiple data sets in solar physics research. With a VSO, the efficiency of these cross-correlative studies would be greatly increased.

With the explosive growth in the Internet and related Information Technologies (IT), it is now possible to link together distributed data archives and analysis software systems into a federated and integrated system – the Virtual Solar Observatory (VSO). This observatory is "virtual" because it exists only in the form of masses of information stored in computer systems. A VSO would greatly leverage the already considerable investment in "glass and metal" observatories by facilitating data mining and increasing the productivity of existing facilities, both space- and ground-based. A VSO would provide users with a common interface to an integrated system of distributed data archives and analysis systems, thereby enabling data exploration and accelerating the scientific process. A VSO would also enable interdisciplinary research, provide an educational tool, and demonstrate the use of advanced IT.

The importance of data archiving and mining to the solar research community was recognized in 1998 by the US National Research Council Space Studies Board Committee on Ground-Based Solar Research (Parker Committee) which recommended:

> "Develop a collaborative NSF and NASA distributed data archive with access through the World Wide Web"

More recently, the National Research Council Astronomy & Astrophysics Survey Committee Report has strongly endorsed the creation of a National Virtual Observatory (NVO) as the top-priority small initiative. The NVO is the night-time astronomical analog of the VSO, and represents the growing interest of the general astronomical community in the possibilities of cross-correlative studies. While the goals of the VSO and the NVO are sim-

ilar, there are substantial differences in the data set characteristics and in the related basic scientific questions that motivate the design concepts. The data set differences are cataloguing requirements, spatial coordinates and transformations, object/pattern recognition needs, and temporal dimensions. The solar scientific community believes that the technical data set differences, warrant the creation of a VSO separate from an NVO.

2. Science-Driven Design Reference Models

The VSO is a solution to the scientific need for analysis of heterogeneous data sets. The community has provided some examples of Design Reference Models that illustrate the range of science progress that would be accelerated by a VSO. One of these model studies, the origin of the solar wind, is described in more detail in Appendix 1 to illustrate some aspects of VSO functionality. Here, we briefly describe several others, along with the types of data that are needed to address the topic.

- Space weather, which seeks to predict geomagnetic storms caused by flares and solar coronal transients such as coronal mass ejections (CMEs), would greatly benefit from more reliable and detailed knowledge of the solar surface signature of an imminent ejection. This requires correlating time series of coronal images, filament positions, and surface quantities such as vector magnetic fields and x-ray emission in several wavelengths.

- The detailed structure of a sunspot atmosphere, where intense magnetic fields can explosively erupt and produce flares and CMEs, is still poorly understood. A composite data set of vector magnetic field maps, space-based UV spectra, and temporal sequences of EUV images would yield a more complete description of the density, magnetic field, flow velocities, turbulence, etc. with height above sunspots.

- The mechanism by which sunspots block solar energy and create fluctuations in the solar irradiance which drives terrestrial weather is poorly understood. Combined observations of the surface vector magnetic field and the spatial intensity distribution in several wavelengths are needed for comparisons with theory.

- The entire uninterrupted life cycle of an active region, which is the location of several sunspots and which can live for several months, has never been observed. Such a study requires multi-day time series of surface magnetic field, doppler velocity, and intensity in several wavelengths from different observatories.

- The statistical evolution of the solar granulation properties over the course of an 11-year solar activity cycle has only been sparsely sampled. A consistently sampled study could reveal clues about the long-term behavior of the underlying near-surface flow field as well as provide insight into the driving of the solar oscillations. Addressing this problem requires multi-year time series of surface magnetic field, doppler velocity, and intensity in several wavelengths.

- The subsurface structure of active regions is beginning to be probed by helioseismology, which uses the acoustic oscillations of the Sun to infer the properties of its interior. Local helioseismology, in which subsurface properties are determined in small patches

rather than globally averaged, must be placed in context. This requires combining surface magnetic field measurements and long time series of doppler velocity measurements. This technique may eventually be able to locate active regions before they emerge on the surface which would be useful for space weather predictions.

- The solar wind, which transports charged particles from the Sun to the terrestrial magnetic field and atmosphere, is believed to originate near the boundaries of coronal holes where the solar magnetic field is open to interplanetary space. However, the details of the physical process that accelerates the wind are still unknown. Research on this question requires images that show the location of coronal holes at either X-ray or infrared wavelengths in conjunction with data on solar wind speed, and surface quantities. A VSO-based study of one possible mechanism is described in more detail in the appendix.

- A long-standing mystery is the mechanism by which the outer solar atmosphere is maintained at a much higher temperature than the solar surface. In addition, it is now known that substantial areas of the atmosphere are actually much cooler than the surface. Unraveling the complex thermodynamic and energy budget of the solar atmosphere requires the spatial and temporal temporal correlation of observations in the infrared, visible, and UV with magnetic and velocity fields at different heights above the solar surface.

Integrated access to a wide variety of solar data sets would thus greatly accelerate progress for many significant solar scientific problems. In the next section we outline a response to this challenge.

3. The Virtual Solar Observatory Concept

In this section, a draft concept of VSO architecture, management, and data sources is outlined. The underlying design drivers are distributed systems, adaptive and scalable expansion, and accountability.

3.1 VSO Architecture

The VSO is a federated system of distributed solar data archives and analysis software. The conceptual model is a single (but mirrored) web service from which a user can carry out a search of all available VSO component archives locate data that fulfill a set of criteria. This will enable the user to efficiently collect the necessary materials for the scientific research or educational project.

A distributed archive is a better solution than the construction of a large centralized data center since the expertise required to fully exploit the data resides with the groups that designed, built, and used the instruments to collect the data. These groups are by far the most qualified to maintain their own data sets. It is clearly not feasible to relocate representatives of each archive to a new physical location.

The VSO should be more than just an integrated tool that searches the static tables of distributed data bases. The VSO will also allow searches based on quantities computed from the data itself in addition to the usual preselected parameters stored in tables. The system will perform the calculations, and present them to the user in graphical form so that

the search can be further refined. Since the computing load of performing complex content-based queries can be substantial, the VSO design will incorporate dynamic load-balanced computing distributed among the available servers at the time of the query.

The user will be provided with analysis and graphical display tools, as both visual programming interfaces, and as downloadable software that can be installed on the user's local machine. It is expected that an object-oriented programming environment will provide the most flexible and scalable solution. This suggests the use of Java-based programming languages for the central system of the VSO.

When ready to order the selected data, the user will be able to select from menus of alternate data formats, for example XML, HDF, or FITS, and data delivery methods.

The VSO must be an adaptive scalable system that grows as additional data sets and analysis packages are added. A major goal of the VSO design philosophy is to lower the level of integration effort facing the administrator of an existing data set since imposing new design standards on existing systems can deter participation and require substantial resources. To this end, the VSO design should be able to bi-directionally translate between local parameter names, programming languages, data formats and data base queries and VSO standards. One possible approach is an intelligent keyword thesaurus to translate local parameter names. The thesaurus would search a list of commonly-used synonyms when confronted with a new parameter table, and update a translation table and synonym list accordingly.

It is essential that the evolving design of the VSO be continually tested and its scientific utility and productivity be validated. To these ends, several evaluation metrics will be incorporated in the VSO design. These metrics will undoubtedly include accounting of the individual local data archive usage, such as data volume, user statistics and CPU loadings. The overall VSO usage will similarly be tracked, as well as the research papers based on VSO-enabled science. If the VSO is an effective solution, then the number of non-VSO accesses to the member data sets should decrease as the number of visits through the VSO increases. Specific target metrics, such as a fraction of published solar experimental papers based on VSO resources, should be established to judge VSO impact.

3.2 VSO Management

The VSO management should reflect the VSO technology – the management should be accountable, distributed, and adaptive.

Perhaps the most important requirement of VSO management is accountability, both to funding agencies, and to the user community. This can be accomplished with a small core management team consisting of a Project Director, Project Manager, and Project Software Engineer. The Project Director would be responsible for the overall success of the VSO, represent the VSO in public forums, provide the point of contact between the VSO, the funding agencies, and the community. The Project Scientist would be responsible for ensuring that the VSO fulfills the requirements of the scientific community, for example, that the VSO contains reliable calibrations. The Project Manager would be responsible for keeping the VSO development on schedule and on budget, as well as operating efficiently. The Project Software Engineer would be responsible for developing the overall VSO Information Technology plan and for verifying the code functionality. The core team would receive frequent (at least weekly) reports on the VSO usage statistics, such as data transfer volumes, CPU loadings, etc.

While this core team provides the project coordination, it is clearly not enough to provide the broad expertise required to make the contents of the VSO useful. This can only be provided by a distributed management structure, such as a Scientific and Technical Oversight Board, with three types of members:

- One representative from each of the component archive institutions. These members, who may be a scientist, IT professional, or manager, will be responsible for representing their institute's data system.
- Members who are primarily users of data, rather than data providers. These members will provide the customer's perspective.
- Members from the Computer Science and Information Technology community, both academic and commercial. They will supply technical input and advise on trends in the computer industry.

Since the VSO is a science- rather than technology-driven, the Oversight Board must be tightly coupled with the core team with, for example, the core team being selected by the Board. In addition, the Project Scientist position could be filled on a rotating basis by board members. This federation of archive managers, users and information technologists is a key to the long-term success of the VSO.

The substantial software effort will be distributed among the participating archives with a full-time programmer at each site. The work schedule of each local programmer will be split evenly between overall VSO software development under the direction of the VSO Project Engineer, and local archive development under the direction of the local archive Project Scientist.

The VSO must continually adapt to evolving scientific, technical, and organizational conditions. The scientific evolution is in the form of new data sources and analysis algorithms. These changes will typically first appear at individual VSO archive components, where they may be developed to become widely-used VSO tools. The technical developments are naturally driven by the IT marketplace, while the organizational changes will likely arise from the addition of member institutions. To adapt to rapidly-changing scientific and technical developments, an Executive Committee, selected from the Oversight Board, should meet at least semi-annually with the core team. The membership of the Oversight Board can be reviewed on an biannual basis. Members for the initial Oversight Board have already been recruited, as listed in Appendix 2.

Since the success of the VSO depends on the health of its components, sufficient resources should be made available to individual components to bring them to approximately the same level of development. This can be efficiently accomplished by administering resources through the VSO management structure. This would allow a close coordination of resource flow with VSO requirements, and provide overall fiscal accountability. It also requires that the VSO core team include a Project Manager.

The VSO management must also allow for interagency cooperation, since the VSO component archives contain data from both NASA and NSF facilities. This requires the reconciliation of some differences in the policies of the agencies. For instance, archives based at NASA centers can only receive NASA resources while other institutions can combine resources. A cooperative agreement between the NSF and NASA in this area would increase

the leveraged return on the scientific investment of both agencies. In addition, international funding agencies should also be accommodated, as there is strong international support for a VSO. In Europe, the Joint Organization for Solar Observations (JOSO) contains a working group (JOSO WG3) charged with developing solar data handling and analysis. The chair of JOSO WG3 has formally expressed interest in the project.

The physical location of the core team remains to be decided.

3.3 Initial VSO Data Sources

While it is anticipated that the VSO will grow as data sets are added, several institutes have already committed to contribute their data sets for participation in an initial VSO. The data sets, currently totaling a volume of 104.3 TB, are from ground- and space-based observations, from the US and from international sources, from NSF- and NASA-funded groups. These include:

- The High Altitude Observatory (HAO) which maintains archives of Mauna Loa Solar Observatory coronal data, Precision Photometric Solar Telescope (PSPT) images, Experiment for Coordinated Helioseismic Observations (ECHO) doppler images, Advanced Stokes Polarimeter (ASP) Stokes profiles, and Chromospheric Helium I Imaging Photometer (CHIP) data.

- The NASA/Goddard Solar Data Analysis Center (SDAC) which maintains archives of data from the OSO-7, SMM, Yohkoh, CGRO BATSE, SOHO, and TRACE missions.

- The Lockheed Martin Solar and Astrophysics Laboratory Archive which contains data from MDI, TRACE, Yohkoh/SXT, and ground-based La Palma observations. These data include images in visible, UV, EUV and soft X-ray wavelengths, and full-disk magnetograms.

- The Solar Physics group at Montana State University maintains data sets from TRACE, Yohkoh/SXT, and the Mees Solar Observatory in Hawaii.

- In Italy, the ARTHEMIS/SOLAR group is a collaboration among observatories in Naples, Trieste, Turin and Florence to develop federated solar archives. These archives include data from the Rome RISE/PSPT telescope, the Italian Panoramic Monochromator (IPM) located at THEMIS in the Canary Islands, and the Trieste Solar Radio System (TSRS).

- The University of Southern California maintains an archive of data from the Mt. Wilson 60-Foot Tower. It primarily consists of a helioseismology data set of dopplergrams and related products.

- The National Solar Observatory has a Digital Library that contains daily full-disk magnetograms, Helium I 10830, Hα and Ca K images, coronal scans, and high-resolution solar spectral atlases. NSO also maintains the GONG helioseismology data set, and will soon begin archiving data from SOLIS which will contain longitudinal and vector magnetograms, filtergrams, and solar spectra.

- Stanford University maintains the SOI Science Support Center, containing the MDI/SOI data from SOHO. This is primarily a helioseismology data set, but it also contains full-disk magnetograms. In addition Stanford archives the magnetic field data from the Wilcox Solar Observatory, and Ca K images from the Taiwan Oscillation Network (TON).

- Big Bear Solar Observatory, operated by the New Jersey Institute for Technology, holds archives of full-disk Hα, Ca K, and white light images.

- After launch, the HESSI mission will create an archive of flare spectra and energetic particle emission data.

All of these groups have expressed interest in linking their data holdings to a VSO. Other data sets that may be added at a later date include NOAA/SEL, the UCLA/Mt. Wilson 150-Foot Tower data set, radio data from Owens Valley, and data from U California/Northridge. The description of existing data sets above is not exhaustive, and further details can be found in Appendix 3.

4. Education and Outreach

Since the Sun is a dominant aspect of everyday life, solar images have long played a role in education. It is recognized among educators that the Sun can be easily incorporated into many parts of the curriculum: astronomy, physics, life sciences, art, literature, and even music (through helioseismology) can all benefit from solar examples. This tradition is still strong today as is evident from the usage of solar educational web sites at Stanford, HAO, and NSO.

The VSO will provide a second gateway into the solar data for non-professional users who do not need highly detailed search and analysis tools. This "educator's interface" will be graphics-intensive, with data selection primarily based on image maps rather than web forms. The VSO will also house complete lesson modules based on solar science.

5. Schedule And Budget

A preliminary estimate, shown in Appendix 4, suggests that the initial VSO development linking 10 archives could be accomplished in two years at a cost of \$4.3M for the first year and \$2.9M for the second year. These costs contain about \$700-800k per year for the core team, and \$225-350k per year per component archive, as detailed in the appendix. One key concept in the budget is that the VSO would provide 1 FTE programmer at each archive for these first two years, of which 0.5 FTE would be required for overall VSO development.

After the initial two-year development, the estimated annual cost of operating a 10-archive VSO is \$1.7M per year. This contains \$585k for the core team, with the 0.25 FTE Project Scientist contributed by the archive whose board member is currently serving in that role. The component archives would be funded at \$115k per year, again with the programmer committed half-time to overall VSO development. With this model, each additional new archive added after the initial phase is estimated to cost an additional \$350k in the first year the archive is added, and \$115k per year subsequently. These costs may well be decreased after the initial start-up and the development of standards for component archive integration.

Appendix 1

A detailed science design model for the VSO
Understanding the effect of acoustic oscillations on the origin of the solar wind

In order to explore the technical issues of VSO construction, consider a small experimental prototype involving five data sets from just two data centers – the NSO Digital Library (NSO-DL) and the Stanford SOI Science Support Center (SSSC). The scientific question is "Is there a significant difference in the energy of the solar acoustic oscillations inside and outside a coronal hole?". This question would contribute to our understanding of the generation of the solar wind.

Note that, while two specific archives are named here for illustrative purposes, they are not the only two that could be used for this project. Other possibilities are SOHO/EIT for coronal hole data, and GONG or TON for oscillation data.

Answering this question requires combining data that provides the coronal hole location as a function of time, with time series of doppler images for the helioseismic analysis. This, in turn, requires querying different data sets to obtain co-temporal data, performing pattern recognition to locate the coronal hole, and then analyzing the acoustic power within and without the hole. It is these types of correlative studies that the VSO would address.

In some more detail the researcher would need to perform the following tasks:

A) Using the NSO-DL, perform a content-based query to locate a time series of He 10830 images containing a candidate coronal hole. The technical requirements are:

1. Pattern recognition of the presence of a coronal hole. Currently this is done by the observer's brain.
2. Parameterization of the hole properties – area, heliographic bounding box, strength.
3. Presentation of the candidate images to the user for further culling.

B) Using the NSO-DL, locate the temporally closest magnetograms. The technical requirements are:

1. Location of magnetograms close to the same time as the He 10830. This is already provided by the NSO-DL.

C) Using the SSSC, locate the corresponding desired MDI data. The technical requirements are:

1. Presentation of the temporal and spatial bounding volume of the NSO-DL observations to the SSSC.
2. Querying of the SSSC to locate data within the bounding volume.
3. Selection among the available data types: e.g. Full-disk dopplergram/magnetogram, Hi-Res dopplergram/magnetogram, etc.
4. Presentation of the candidate images to the user for further culling.

D) Take delivery of all relevant data. The technical requirements are:

1. Provision of the user's location.
2. Selection of a data format and delivery method.
3. Implementation of the delivery.
4. NOTE: The SSSC has an automatic system for this already.

E) Create acoustic power maps inside and outside the coronal hole. The technical requirements are:

1. Specification of several new spatial and temporal bounding volumes for the acoustic power maps.
2. Extraction of the data volumes.
3. Computation of the maps – requires specification of the remapping, and the frequency and wavenumber bands.

F) Analyze and display the results. The technical requirements are:

1. Overlaid translucent color graphical displays of the power maps and co-located portions of the He 10830 images and magnetograms.
2. Line plots of acoustic power, He intensity, B.
3. Scatter plots of acoustic power vs He intensity, B, etc.
4. Statistical analysis – regression, significance, etc.

Appendix 2

Potential initial members of the VSO Scientific and Technical Oversight Board

- University of Colorado - Thomas Ayres
- Stanford University - Rick Bogart or Jim Aloise
- High Altitude Observatory - Peter Fox
- NASA/Goddard Solar Data and Analysis Center - Joe Gurman or Luis Sanchez
- National Solar Observatory/GONG/SOLIS - Frank Hill
- Lockheed/TRACE - Neal Hurlburt
- Montana State University - Piet Martens or Alisdair Davey
- ARTHEMIS/SOLAR - Kevin Reardon or Mauro Messerotti
- University of Southern California - Edward Rhodes
- SOHO/CDS - William Thompson
- Big Bear Solar Observatory/New Jersey Institute of Technology - John Varsik
- HESSI - Dominic Zarro

Appendix 3

Some details of Initial VSO Component Archives

High Altitude Observatory - contributed by Peter Fox

The majority of all solar raw and processed data from MLSO (except some early material which is still on film) and other HAO observing sites (images and spectra) are archived on the NCAR Mass Storage System (MSS). For some instruments, additional backups are retained on magnetic tape (variety of media formats). The total archive volume is 5 TB. An incomplete list of the holdings is below.

Current processed data and analysis products (e.g. overlays, movies) for a variety of instruments are also maintained on-line within HAO, and available via http or ftp. File services are performed on Sun Sparc and Ultra servers running Solaris, Apache web servers, and DODS middleware. Access ranges from anonymous ftp all the way to a fully integrated searchable catalog and data retrieval via the web. Community use of quantitative data ranges from 3-4 users for small instrument projects to more than 50 (cumulative) for coronal and chromospheric images. Qualitative use (web page hits) is high, e.g. 25,000 hits/month of just the MLSO home page and 'images of the day'.

HAO has an existing data services plan and is in stage two of four stages of implementation. The latter two stages consist of 1) carrying the prototype implementation to all HAO data services in the form of a uniform method of searching and retrieving data and allowing access via commonly used analysis tools, and 2) taking advantage of http access to NCAR's MSS for access to all archived data. Until 2) is implemented, only on-line data is available via 1). A DVD-jukebox (the new 9 GB disks, 100+ disk capacity) is being considered as an alternative.

Incomplete list of HAO data sources:

- Mark IV white light coronameter (MKIV) - polarization brightness, MLSO –every 3 minutes.
- Chromospheric Helium I Imaging Photometer (CHIP) - about He I 10830, MLSO – wavelength scans every 3 minutes.
- Polarimeter for Inner Coronal Studies (PICS) and low-K corona - Hα, MLSO – on disk in three wavelengths, every 3 minutes. – off disk in three wavelengths, every 3 minutes.
- Experiment for Coordinated Helioseismic Observations (ECHO), MLSO/Tenerife. – red/blue wing of Potassium line, for doppler images.
- Precision Solar Photometric Telescope (PSPT), MLSO (data from NSO/SP will also be available) – Ca K, red and blue continuum up to every 5 minutes.
- ASP - Advanced Stokes Polarimeter - Stokes profiles, by request.
- SPARTAN
- STARE

- Mark III coronameter
- Digital Prominence Monitor
- Prominence Monitor
- SMM/Coronagraph
- Skylab/Coronagraph
- RISE Spectral Synthesis

NASA/Goddard Solar Data and Analysis Center - contributed by Joe Gurman

Types of data: images, spectra, and time series of single-point (e.g. irradiance) data from OSO-7, SMM, Yohkoh, CGRO BATSE, SOHO, and TRACE. Solar data from the HESSI and STEREO missions are expected in the future, and likely data from other solar missions as well.

Media: the older missions' data are on 4 mm DAT tape (roughly 300 tapes here or at an offsite storage facility), the Yohkoh data, which started on DAT, are now on CD-ROM, and will someday be placed online.

All SOHO and TRACE data kept here (but no MDI high-rate data from SOHO, since those are archived at Stanford) are kept on network-attached file servers (i.e., no interactive OS), and are available through Web-based interfaces, with the final, compressed files being deposited on an anonymous ftp drop box for download.

Total volume of online storage: 3.3 TB.

Lockheed Martin Solar and Astrophysics Laboratory Archive - contributed by Neal Hurlburt

MDI, TRACE & Yohkoh/SXT, LaPalma archive

Data: Images in visible, UV, EUV and Soft X-ray; full-disk magnetograms.

Current storage (by archive):

Archive	Disk	Tape	CDROM	Server Access
TRACE	400 GB	400 GB	0	SGI/IRIX Web/SolarSoft
MDI	30 GB	0	0	SGI/IRIX Local/SolarSoft
SXT	9 GB	0	500 GB	SGI/IRIX Web/SolarSoft
LaPalma	30 GB	3000 GB	0	SGI/IRIX Local/SolarSoft

Downloads: TRACE: 10.4 GB/month 15k Images/month from over 100 IP addresses

Montana State University - contributed by Piet Martens

Three primary data sources are in use at MSU: Yohkoh SXT, TRACE, and Mees Spectral Imaging data.

SXT data is primarily in the form of CD-ROMS stored in a 500-CD Jukebox and amounts to 300 GB. There is also 240 GB of disk space in use for the SXT movie reformatting projects.

An archive of TRACE data which currently occupies 215 GB is maintained.

The Mees data is mostly stored in the form of 8-mm Exabyte tapes some of which are being converted to CD-ROM. This is just over 8 TB of data

Hardware: Pioneer CD Jukebox is served via NFS to solargroup machines. The controlling software is on a Solaris machine. The TRACE archive is maintained on an SGI using software RAID 0. The SXT movie reformatting project uses a hardware RAID 5 configuration based on a Linux/x86 box.

Data access is normally via local collaboration for which online accounts are provided. Other data is furnished by anonymous FTP.

ARTHEMIS/SOLAR – contributed by Kevin Reardon & Mauro Messerotti

ARTHEMIS is a project of the Capodimonte Astronomical Observatory (Naples) that uses modern relational database technologies to facilitate and improve the archiving the solar data.

We currently archive the data from the RISE/PSPT telescope at Rome and from the THEMIS solar telescope on Tenerife in the Canary Islands. The RISE/PSPT data consist of daily full disk images of the Sun obtained in Ca K, blue continuum, and red continuum. The full archive of images obtained by this instrument since its installation in 1996, occupying approximately 8 GB, is available on-line using an intuitive search interface.

The data from THEMIS come from the Italian Panoramic Monochromator installed on that telescope. During the 1999 observing year, the instrument obtained 25 GB of data. For the year 2000 the data flux is expected to be 100 GB. A future generation of the instrument to be installed in 2002 will produce several TB per year. THEMIS will also start obtaining full disk images, dopplergrams, and magnetograms, which will also be archived by ARTHEMIS and available from our web pages. The data will be stored in a DLT jukebox allowing rapid access to a large subset of the data. The acquisition of a writable DVD jukebox is also being considered.

The ARTHEMIS archives uses an Oracle RDBMS running on a Compaq/Digital AlphaServer currently with 35 GB of disk space (to be expanded). The web interfaces for the end users are provided using the Oracle PL/SQL programming language that allows direct interaction with the data stored in the database tables. We also use this programming language to provide value-added services, such as an interactive solar ephemeris. We currently receive approximately 3000 web page hits per month at the archive at the URL `http://arthemis.na.astro.it/`.

SOLAR is a collaboration among the astronomical observatories of Turin, Capodimonte, Trieste, and Florence to produce a combined solar data archive. Most importantly, SOLAR will be one of the three ESA-approved European nodes of the complete SOHO data archive. The SOHO archive's current size is more than 500 GB (with another 1.5 TB still to be processed and included) and also includes related ground-based synoptic data and data analysis software. The project also envisions linking SOLAR, ARTHEMIS, and other online databases using distributed database technologies. These other databases include:

- SOLRA, discussed below
- data from the Catania Astronomical Observatory which obtains daily full-disk images of the sun in H-alpha and the continuum.

The SOLAR archive is being designed to have a long lifetime – a minimum 10 years after the end of the SOHO mission – and is expected to be completed by the end of 2001.

SOLRA, the SOLar Radio Archive, is a SOHO-compliant radio data archive currently under development at the Trieste Observatory. SOLRA will provide access to radio data from the Trieste Solar Radio System (TSRS) through a WWW interface and will be integrated in the European SOHO archive under development in Turin. The TSRS is a unique, dedicated solar radio facility presently active in Italy for the radio surveillance of the solar corona. The TSRS instrumentation is a set of two multichannel solar radio polarimeters operated continuously on a daily basis. The Metric Multichannel Solar Radio Polarimeter (mMSRP) has a 10 m antenna and 4 receiving channels at 237, 327, 408 and 610 MHz, whereas the Decimetric Multichannel Solar Radio Polarimeter (dmMSRP) has a 3 m antenna and 2 receiving channels at 1420 and 2695 MHz. Both instruments record the time evolution of the radio flux density and circular polarization with high time resolution (1 millisecond for routine synoptic observations and up to 0.1 ms during coordinated campaigns). In a short time, radio indices averaged on 10 minute time intervals will be published on the Internet in near-real-time in the frame of the Italian Space Weather Initiative. This initiative is based on the solar and solar-terrestrial observing facilities operating in Italy, as a precursor to a European network.

University of Southern California/Mt. Wilson 60-Foot Tower - contributed by Ed Rhodes

Within the last several months a concerted effort to catalog the wealth of data obtained at the 60-Foot Solar Tower at the Mt. Wilson Observatory has begun. This cataloging includes the creation of a complete listing of all of the available data products that have been obtained and processed since 1987, as well as listing of the daily solar images which have been posted to the URL `http://physics.usc.edu/solar` since June of 1999. The types of images which are now posted on this website include daily dopplergrams, daily magnetograms, and filtergrams which are taken both in the morning and afternoon of each clear day.

The archive listing of all of the available data products is being assembled into HTML format and will eventually be available from the 60-foot Solar Tower web site. Because of the small size of the group, this process is slow, but this feature should be online sometime during the summer of 2000. Because of the size of the raw observational database and the additional base of reduced data products, both of which have been assembled over the years since 1987, it is not possible to store all or even a substantial portion of these data online. The raw data is written directly to Exabyte cartridges and can exceed 2.5 GB per day during long summer days. While each of the reduced data products takes a diminishing amount of storage space with each subsequent step in the processing pipeline, the total amount of available data is enormous.

To develop an estimate of the amount of raw and processed data in our archives we have assumed an average run duration of 9 hours per day. Factoring in an estimate of the annual number of days lost to equipment problems and cloudy weather, which varies somewhat from year to year, we arrive at a rough estimate of 600 to 700 GB of raw data that is collected per year.

Taking into account only the last decade of observations and processing, we then obtain the following archive estimates:

Raw data (1024x1024 pixel Na D filtergrams)	5100 GB
Uncalibrated 1024x1024 pixel dopplergrams	870 GB
Sets of Spherical Harmonics Coefficients	402 GB
Transposed Time Series of S.H. Coefficients	197 GB
Power Spectra	
1990 observations	120 GB
1996 observations	70 GB
1997 observations	244 GB
Total of Raw and Processed Data Products	6.6 TB

All of this data is stored on several thousand 8mm Exabyte data cartridges. We have available a total of seven Exabyte tape drives, four of which are housed in two Exabyte EXB 210 jukebox systems.

The current website location is being served from the Sun Enterprise 3000 server of the USC Department of Physics and Astronomy. The total amount of online disk storage space which is currently available for the 60-Foot Tower webpage is under 2 GB. Even this amount of available disk space is under daily pressure from our ongoing data reduction tasks as the total amount of online disk storage space available to our entire group on this server is only about 6 GB and our data reduction storage needs are constantly growing. At the moment, our webmasters have to limit the number of images in the current database by removing older images to accommodate each set of new images. While this is not desirable, it is necessary to handle our current production rate of 16MB of new images per day, while at the same time maintaining a reasonable number of days of data in the online area.

Currently, the daily set of images at `http://physics.usc.edu/solar` constitute a filtergram, dopplergram, and magnetogram taken in sodium light. However, we also soon hope to add three additional daily images taken with a Potassium version of our MOF, a series of temporally filtered dopplergram(s), magnetogram movies, and possibly even some flare movies.

While the 60-Foot Solar web pages have been available on the internet for some time now, daily images have only been available for a little under one year. The number of hits to the website grew dramatically since June 1999 when the image database became available online. A year ago, we had only 300 hits per month, now we keep climbing steadily to over 1000 monthly hits since the end of March.

A more efficient arrangement for the portion of out database which is available online at any given moment would be to remove it from the current Departmental server and to replace that server with a dedicated unix-based server which would only house both the web pages and the temporary online data subset of our total archive. I would envision a single-processor Compaq Alpha 21264 600 MHz system with an attached disk farm of 200-300 GB.

National Solar Observatory - contributed by Frank Hill

NSO currently holds two major data sets, the NSO Digital Library and the GONG data set, and will be adding the SOLIS data set in the near future.

The NSO Digital Library currently contains about 200 GB of solar data distributed over magnetic disks and a CD-ROM jukebox system. The data holdings comprise:

A) the entire Kitt Peak Vacuum Telescope (KPVT) synoptic data set of full-disk solar magnetograms and Helium 10830 spectroheliograms from 1974 to the present

B) the entire Fourier Transform Spectrometer (FTS) data set of high-resolution spectra and interferograms of solar, stellar, planetary, terrestrial atmospheric, and laboratory sources

C) recent full-disk solar spectroheliograms in Ca K and Hα from the NSO/Sac Peak Evans facility.

All of these data are on-line and accessible through a web-browser user interface to a searchable relational data base system (RDBMS). The URL is `http://www.nso.noao.edu/diglib`.

NSO Digital Library Usage statistics for 1 July 1998 - 30 June 1999:

2095 FTP users

12,752 FTP logins

24,067 files downloaded via anonymous FTP

115,390 web page hits

Demographics for that period:

	FTP Users	FTP Logins	FTP file downloads
US Science	17%	48%	43%
US Public	41%	17%	7%
Foreign	42%	36%	49%

Distribution of downloaded data products that period:

71% KPVT (magnetograms, synoptic maps, Helium images)

20% FTS (Spectral atlases, general archive)

9% Sac Peak Spectroheliograms (Hα, Ca K images)

The GONG data set comprises approximately 6 TB of helioseismic data in the form of full-disk dopplergrams, intensity, and modulation images obtained every minute at 6 sites; spherical harmonic time series and acoustic oscillation power spectra; and full-disk magnetograms every twenty minutes. In FY2000, a higher-resolution camera will be installed raising the ingest rate of the GONG archive from 1 TB to about 20 TB per year. Currently virtually all of the data is off-line on magnetic tape media.

SOLIS will have a major impact on the NSO Digital Library. The core SOLIS data products will include 3-per-day photospheric vector magnetograms, chromospheric line-of-sight magnetograms, and He I 10830 quantities; 1-per-minute Hα, He I 10830, continuum and Ca K images; and 60 disk-integrated solar spectra per day. The core science programs will produce data at the rate of over 30 GB per day, and are expected to total nearly 100 TB over the 25-year lifetime of the instrument. In addition, it is conceivable that some intensive PI programs could operate the full suite of instruments at their maximum rate, which would produce 240 GB in a single day. These prodigious rates, along with the requirement to allow data access over the Internet within 10 minutes of collection, cannot be handled without

substantial hardware upgrades both to the data link connecting Kitt Peak with downtown, and the NSO Digital Library. The data link must have a DS-3 capacity to handle the maximum volume of SOLIS data production, and a proposal to the NSF to fund this has been submitted. The Digital Library hardware must also be substantially upgraded, and we anticipate installing a mixture of optical disk (probably DVD jukeboxes), magnetic disk (large RAIDs), and high-capacity tape (DLT jukeboxes) systems to at least partially handle the storage of the core science data.

Stanford University - contributed by Rick Bogart

The principal data archives at this site are the complete and ongoing data sets from the SOI-MDI mission on SOHO (1996 - present) and from the Wilcox Solar Observatory (WSO) (1975 - present). The catalogs and data archive and distribution services for these efforts are distinct and not integrated. In addition, the Taiwan Oscillations Network (TON) and the Mt. Wilson 60-foot Telescope make use of the SOI data archive infrastructure to distribute their data. Archiving of the TON data is proceeding very slowly and laboriously due to media problems, but a low-level effort continues. The SOI-MDI and WSO data archives are complete and keep pace with the flow of new data. Note that these are the data archives of the Solar Observatories Group only; there may be other solar data archives in the hands of researchers at Stanford University with which we have no involvement.

The SOI-MDI data archive consists principally of photospheric dopplergrams, white-light photograms, and magnetograms at various spatial resolutions covering the full-disc and an area of 10% or less near disc center at higher resolution; there are also some line-depth images and filtergrams. The principal observing cadence is one image per minute, but there are some collections at both higher (2/min) and lower (0.2 - 0.083/min) cadences. In addition the archive contains substantial quantities of data products derived from the image time series, including means, synoptic maps, mode amplitudes and frequencies, and power spectra. The total volume of the data archive is currently 66 TB, of which about two-thirds is valid data from the project. Virtually all of the data reside offline on Ampex tape; there is an online disk cache for data staging of about 1 TB. The archive is served principally by a collection of SGI machines running IRIX. Data are accessible via a web interface for distribution via ftp, rcp, and on offline media (Exabyte, DAT, CD-ROM, and Ampex tapes). Since the beginning of the SOHO mission (end of 1995) we have serviced approximately 8500 data requests from around 400 different users for about 110,000 individual data sets comprising a volume of roughly 14 TB.

The data archiving agreements with TON and Mt. Wilson contemplate our archiving comparable types of image data from those observatories and producing and archiving the same kinds of derived data products. So far, only a handful of uncalibrated Ca K photograms from the TON are incorporated.

The WSO data archive consists of daily low-resolution full-disc magnetograms and several (10) mean-field measurements per day.

There is no practical way of placing the full SOI data archive online; the WSO archive is online. Web-based interfaces already exist.

Big Bear Solar Observatory - contributed by John Varsik

BBSO has been undergoing major changes in instrumentation over the last two years. Our set of telescopes allows us to operate six high-resolution instruments at various wavelengths

plus an Hα full disk camera simultaneously. Before 1998 most of these instruments recorded to either videotape or film. Even so, the archive currently contains 3.4 TB of digital data, mainly magnetograms and Hα filtergrams, with some magnetograph data going back to 1986.

Currently almost all data is off-line. BBSO does have an FTP server, on which daily images from each instrument is posted. These generally include full-disk Hα, Ca K, and white light images, as well as images from the high-resolution cameras (usually Hα, magnetograms, and Ca K) that are running that day. These images are stored in JPEG format on the FTP server, going back to at least 1998, and sometimes earlier depending on the instrument. The current FTP server is a Pentium 166 MHz machine running Linux. The disk is about 4 GB in size, although it will soon be replaced with a 36 GB disk. Access is through anonymous FTP, with a link to the archive through the Web server. Although there are no statistics for the FTP server itself, the Web server shows access over the last month from nearly 24000 unique Internet sites, with over 66000 total visits.

The recent replacement of the video cameras with digital cameras has greatly increased the maximum rate of data acquisition to more than 10 GB per day. While this rate will not be reached every day, the Hα full disk system (with a 2032 $\times$ 2032 CCD) alone acquires 5 GB every clear day.

Several upgrades are planned for the on-line archive system in the near future. In addition to the disk upgrade, JPEG versions of all of our images will be placed on the FTP archive for on-line access. A simple Web-based search engine to allow searches of the on-line data will also be implemented. The lifetime of the on-line data will be determined by available resources. At this point the JPEG data on-line lifetime is expected to be at least one month.

As always, FITS-format copies of the original data are available upon request, although the availability of this is limited by staff time and other resources.

Obviously keeping data on-line for a longer time or in the original FITS format would require substantially greater resources than are currently available. Our camera systems are capable of generating more than 10 GB of data per day, depending on the experiments that are running. While not all of this would ever be on-line in its original form. JPEG compression allows the storage of (sometimes in reduced resolution) compressed images of all data from one day onto a single CD-ROM. In addition, BBSO is collaborating with observatories in Austria and China to establish a 24-hour Hα full disk monitor. This could in principle generate another 10 GB of original data per day.

Upgrading the system to keep a year's worth of compressed data on line would require up to a 500 GB disk array or a CD-ROM jukebox, with an appropriate computer system to run it. Such a computer system might cost $5000 for each year of data to keep on-line. Additional years would scale accordingly. Note that this is for the compressed data, not the original FITS files.

A long-term goal would be to digitize some of the material in the photographic and video BBSO archive, which goes back to 1970. Significant resources would be required.

Appendix 4
Estimated VSO Budget and Schedule

Year 1: Set up VSO Core Team and Server, begin development at component archives

Year 1 Budget
Core Team

Item	Cost (k$)
Project Director	150
Project Manager	100
Project Engineer	125
Project Scientist	100
Administrative Assistant	60
VSO Server	150
Overhead	100
Total 1st-Year Core	785

Note: All positions 1 FTE, server is Multi-CPU, with large RAM and RAID.

At each component archive

Item	Cost (k$)
1 FTE Archivist	50
1 FTE Programmer	125
0.25 FTE Scientist	25
Server/Storage	150
Total 1st-Year per archive	350

Note: Programmer will work 50% time on overall VSO, 50% on local archive.

With 10 archives: 1st year total: $10 \times \$350\text{k} + \$785\text{k} = \$4285\text{k}$

Year 2: As Year 1 but equipment costs fall to \$50k for the Core team, \$25k per archive.

With 10 archives: 2nd year total: $10 \times \$225\text{k} + \$685\text{k} = \$2935\text{k}$

After start-up:

Maintenance Year Budget
Core Team

Item	Cost (k$)
Project Director	150
Project Manager	100
Project Engineer	125
Project Scientist	0
Administrative Assistant	60
VSO Server	50
Overhead	100
Total Core Yearly Maintenance	585

Note: Project Scientist now 0.25 FTE, donated by local archives on rotating basis.

At each component archive

Item	Cost (k$)
1 FTE Archivist	50
0.25 FTE Programmer	30
0.10 FTE Scientist	10
Server/Storage	25
Total Yearly Maintenance per archive	115

Note: Programmer will work 50% time on overall VSO, 50% on local archive.

With 10 archives: yearly maintenance total: $10 \times \$115\text{k} + \$585\text{k} = \$1735\text{k/year}$

Adding an archive later would cost an additional \$350k for the first year the archive is added, and an additional \$115k/year thereafter