

Steps towards a virtual solar observatory

A. Csillaghy,
Space Sciences Laboratory, University of California, Berkeley, CA 94720-7450; csillag@ssl.berkeley.edu

D. M. Zarro,
*Emergent Information Technologies, Inc., 2600 Park Tower Drive, Vienna, VA 22180;
dzarro@solar.stanford.edu.*

S. L. Freeland,
Lockheed Martin Solar & Astrophysics Lab, Palo Alto, CA 94301; freeland@penumbra.nascom.nasa.gov

Overview

The sun is the most observed star. Hundreds of telescopes observe the multitude of phenomena associated with its activity, such as flares, coronal mass ejections, oscillations, etc. The proximity of the sun gives us the opportunity to observe these phenomena with relatively high spatial resolution. This not only helps investigating phenomena taking place on far more remote stars, but also helps understanding the local space conditions in which the Earth evolves.

Often, observations are coordinated in campaigns involving several instruments. Coordinated observations focus on specific targets—e.g. solar “active” regions (Canfield, 1999). During coordinated observations, the target is observed in a broad range of the electromagnetic spectrum, each individual instrument focussing on the observation of a specific, relatively narrower range. Spacecraft participating in coordinated observations include the High Energy Solar Spectroscopic Imager (HESSI, X-ray and Gamma-rays imager and spectrometer); instruments of the Solar and Heliospheric Observatory (SOHO, spectrometer, chronograph, Doppler imaging, etc.); the Transition Region and Coronal Explorer (TRACE, extreme UV and UV); the Japanese soft X-ray and hard X-ray imager Yohkoh. On the ground, over 45 instruments also participate in coordinated observations. This includes the Big Bear Solar Observatory (H alpha imager), the Owens Valley Solar Array (OVSA, radio imager), the Synoptic Optical Long-term Investigations of the Sun program in Arizona (SOLIS, vector spectromagnetograph, full disk patrol, spectrometer), and Phoenix-2 in Zurich (radio spectrometer).

Coordinated observations generate huge amounts of data. Each individual instrument may record several gigabytes (Gbytes) per day. For instance HESSI, which will be launched in spring 2001, will deliver about 1 Gbyte of raw data daily. In addition, ancillary and derived data are often added to the raw, “primary” data. In the case of HESSI, about 300 megabytes (Mbytes) of quicklook data and tables will be added to the original observations.

Solar physicists want to extract relevant information from these data sets. To do this efficiently and effectively, they need to be supported by software systems helping them to *retrieve* and *analyze* the data corresponding to the relevant information. For solar physicists (and for any interested user in general), retrieval and analysis tasks should be as simple as possible. In other words, the software support should be as transparent as possible. However, in practice, they are everything but transparent. In general, users need to understand the technical details of each individual instrument before they can start working with the

data. The lack of transparency of software systems is not only an issue in solar physics. It is an issue in most scientific fields dealing with large amounts of data.

There are two major challenges in making software systems more transparent. First, data *retrieval* should be globally organized. A single query engine should be able to locate and retrieve, over the Internet, any data sets that correspond to a given specification, independently of their origin. Second, data *analysis* should provide access functions common to any instrument, encapsulating technical details behind a simple standard interface.

In this article, we first present three categories of sites from which solar data can be retrieved. We describe the retrieval software systems associated with these sites. We then present software systems used for analyzing solar data. We show that both retrieval and analysis systems provide only limited global organization and/or standard access of data. We thus also discuss extensions to these systems, including data combination (or mapping) software, re-usable software components, and visualization software. We argue that such extensions could eventually provide the level of software support expected by users.

Retrieval of solar data

In general, solar observations can be retrieved from one of these three site categories:

1. Primary data archives provide access to raw data of individual instruments;
2. Synoptic data archives provide access to summarized and integrated data from multiple instruments;
3. Scientific data warehouses provide access to the same data as synoptic data archives and to derived data.

These categories are obviously related. Synoptic data archives retrieve data sets from primary data archives, summarize them, and store them into instrument-independent data formats. The raw data is *integrated* into the synoptic archive. Data warehouses retrieve data sets from both primary and synoptic data sites. They integrate them, and generate derived data to support on-line data analysis operations (see below).

Before analyzing data from several instruments, users need to retrieve data from each site individually. At each site, they have to enter equivalent data selection parameters to retrieve the corresponding data.

We now describe the advantages and limitations of each site category

Primary data archives

Primary data archives store raw data and catalog information to access them. Raw data can have many different types. Common types are images and tables, but also other formats are used to store spectrograms, audio, or video files. In general, data are stored in binary formats, unique to each specific site, that require a site to provide special reading software. A better alternative, used by an increasing number of sites, is to use a self-describing standard file format, such as the Flexible Image Transport System FITS (Wells et al., 1981). Generic FITS reading software is available, which allows reading the raw data on virtually any computer.

Retrieval systems of primary data archives differ significantly in the flexibility they offer for selecting data sets. The simplest retrieval system associates a URL address with the *file system* storing the data. The file

system is organized chronologically; i.e. files are divided into directories that correspond to a date and time. Relevant files are selected by browsing through the directories of the file system.

However, access to the file system alone does not allow using general search criteria, such as flare class or active region number (two search keys for solar data). Therefore, some primary archive sites offer more flexible systems involving a *relational database* management system and programs called middleware, which mediate between the database system and the web server. This approach requires developing the middleware, and inserting the catalog of observations into the database system.

URL addresses of a few primary archive sites are listed in Table 1.

Table 1: A few primary data archive sites for solar observations

Name of the site	Address
Big Bear Solar Observatory	bbso.njit.edu
HESSI	hessi.ssl.berkeley.edu
Ovens Valley Solar Array	www.ovsa.njit.edu
SOHO	sohowww.nascom.nasa.gov
SOLIS	www.nso.noao.edu/solis
TRACE	vestige.lmsal.com/TRACE/
Yohkoh	www.lmsal.com/sxt
Zurich radio spectrograms	www.astro.phys.ethz.ch/rapp

The Zurich primary data archive

The Zurich solar radio data archive is an example of a primary data site. This site stores data from daily observations of the solar radio emission. The archive contains data recorded over the past 30 years by three radio spectrometers (Perrenoud, 1982, Benz et al, 1991, Messmer et al, 1998). Observations are visualized as spectrograms, where the elements of a 2-dimensional array are associated with the radio flux at a specific time and frequency. Zurich data are stored in FITS format (Csillaghy, 1997).

Spectrograms from the Zurich site can be retrieved with two retrieval systems. The first system allows browsing through chronologically ordered directories. The second system, called ASPECT, allows more elaborate queries (Csillaghy and Benz, 1999). Users of ASPECT interact with the system by switching back and forth between three modes: a query mode, a browse mode and an inspection mode. The query mode allows entering preliminary information about the searched images, such as the type of radio emission searched. The browse mode displays a summary list with image icons. The inspection mode displays summary information about individual data sets, and allows downloading the raw data.

ASPECT has also a data mining capability. From the inspection mode (in which data can be downloaded), one can search for similar images. This allows for the finding of other data sets that could not be found with queries based only on date and time. However, in this type of retrieval, defining the similarity criteria is challenging. Also, generating the parameters necessary for locating the similar data sets (called *indexing features*) can take a significant amount of processing time.

Synoptic archives

Synoptic archives store data from multiple primary data sites. Therefore, they allow retrieving related data sets from several instruments in a single query. Data in synoptic archives are stored in standard formats (for example FITS) that usually comply with rules valid for a set of similar instruments (Freeland, 1999). Raw data from single observatories are periodically fetched, cataloged and, if necessary, transformed into the standard format.

Retrieval systems for synoptic data archives allow searching using the parameters common to several instruments. Users must enter these search parameters only once. For instance, specifying a date and time of observation will return a list of data from several observatories that have been recorded at that time. The data sets selected can be retrieved as original FITS files for analysis, or as GIF/JPEG images for display. Several synoptic data archives are listed in Table 2.

Table 2: Synoptic data archive sites for solar observations

Name of the instrument	Address
Max Millennium Flare/Synoptic Archive	orpheus.nascom.nasa.gov/~zarro/gbo/
SOHO Synoptic Archive	sohowww.nascom.nasa.gov/cgi-bin/synop_query_form
Solar Data Analysis Center	umbra.nascom.nasa.gov

Unfortunately, it is not possible to “manipulate” the original data from the synoptic archive site. Once data sets are integrated into a synoptic archive, they become independent from the original data sets. As a result, a change in the data at the primary site may imply the replacement of all the data at the synoptic archive, to avoid inconsistencies between those sites.

SOHO Synoptic Data Archive

The SOHO synoptic data archive is an example of online-accessible data archive (sohowww.nascom.nasa.gov/synoptic). It allows retrieving space- and ground-based observations used for coordinated planning and analysis of SOHO observations. The SOHO synoptic data archive is updated daily with links to available images that have been copied from remote sites.

Max Millennium Flare/Synoptic Data Archive

The Max Millennium Flare/Synoptic data archive enhances the SOHO synoptic data archive (Zarro et al., 1999). It includes additional observations of solar flares. It provides a central archive of ground-based and space-based data sets for convenient joint analysis with HESSI data. The Flare/Synoptic data archive will eventually include data sets such as images, spectra, light curves, spectroheliograms, that were obtained at relevant times during HESSI-observed events. The data sets are obtained with the prior approval of each data source provider, and when possible, saved locally in standard formats (FITS).

Data from the Flare/Synoptic archive is retrieved using a single web interface. The panel below shows a graphical user interface that demonstrates coordinated retrieval of HESSI events and associated synoptic observations. Searching is based only on date/time ranges. The synoptic data archive includes selected

Yohkoh, SOHO, and ground-based imaging observations. Selection of a HESSI event number returns a list of locally stored flare/synoptic data sets for the current day. Selection of one of these data sets returns a quicklook display of the corresponding data.

Figure 1: The Flare/Synoptic data archive search interface.

The basic synoptic archive described above can be extended to include additional new data types. Finally, it is anticipated that each unique HESSI event will be assigned a numeric event number that will allow localizing the corresponding data into the HESSI catalog. Therefore, the Flare/Synoptic data archive will enable rapid and efficient cross-referencing and searching for overlapping (temporal and spatial) HESSI and synoptic datasets.

Scientific data warehouses

Data warehouses store derived data, which complement the data from synoptic archives. Derived data consist of pre-processed data that are likely to be used in subsequent on-line data analysis tasks. They allow increasing the efficiency and effectiveness of data retrieval from the warehouse.

Togther with derived data, scientific data warehouses provide on-line data analysis services. On-line data analysis mixes data analysis tasks with data retrieval functions. Parameters are not only search keys, but also control parameters that activate processing operations. In this approach, it is assumed that different users repeatedly redo many equivalent operations. These operations can be processed the first time they are requested, and the data product generated can be reused for any further equivalent requests.

Warehouses provide also more elaborate retrieval possibilities than synoptic data archives. They rely on an extended database schema. This allows specifying more search keys, thus queries can be more precise, and finding relevant data sets can be faster.

Warehouses also rely on data copied, summarized, and transformed from primary or synoptic data sites. Therefore, a change occurring in the original parameters (e.g. calibration) in a primary data site for instance will not be reflected directly in the data products available at the data warehouse. Moreover, such a change would require the data warehouse to update all pre-processed data products, which may be extremely time consuming.

HESSI Experimental Data Center

The HESSI Experimental Data Center (HEDC) is an example of a scientific data warehouse for solar observations (www.hedc.ethz.ch). HEDC will complement the HESSI original catalogs with an extended metadata catalog that will support users in selecting relevant data. HEDC will also provide a number of on-line data analysis tasks and store the data produced by these operations for subsequent uses, as described above. This experimental data center is planned to be operational at the start of the HESSI mission.

Analysis of solar data

The goal of data analysis is to extract relevant information from the data. In solar physics, data analysis tasks include operations such as time series analysis, spectral fitting, visualization, solar limb fitting, grid overlay, coordinate transformations, noise reduction, zooming, etc.

In principle, data analysis is straightforward: it “only” consists of a series of selections and transformations. A specific input data set is selected from a (usually large) data archive, and then transformed into another data set by a given algorithm. The way the algorithm works depends on control parameters that are given by the user at run-time. However, most of the time data analysis means dealing with instrumental (i.e. technical) characteristics. This costs a significant amount of time and effort. For each instrument, a specific way of analyzing data needs to be learned.

Software provided by primary data archive sites

Traditionally, data analysis software systems are instrument-specific. One of the reasons for this limitation may be that budgets are associated with instruments, and developing generic software can be considered outside of that budget. In general, the development of a data analysis software system is associated with the development of an instrument. Furthermore, instrument teams are specialized in the operations of a single instrument, and may not have the expertise in developing software in a broader context.

Nevertheless, there is a trend to develop generic software systems. Project investigators are aware that the development of data analysis software is expensive. Re-using software parts that have been developed for another instrument can spare an expensive development time (and avoid reinventing the wheel). Moreover, the newly developed systems are more robust, because the re-used parts of the software have already been tested. Finally, software re-use gives the possibility for small institutions to develop programs that they could not afford to develop alone. As a result, the scientific return of individual projects can be significantly enhanced.

We now describe the limitations and advantages of several data analysis software systems.

The Ragview system

Ragview is an example of a data analysis software system provided by a primary data site (Csillaghy, 1998). It contains tools to analyze solar radio spectrogram data, and a graphical menu interface to work with these tools. It has been designed and developed for the data from the Zurich site presented above.

The system is divided into several modules. A file-handling module provides methods to write and read data. A modification module gives a list of algorithms to handle background, to interpolate, fit models, etc. A display module allows to plot, zoom, and print data. A customization module allows configuring the settings for specific data analysis tasks.

Ragview has been integrated in the Solarsoft system (see below) and is currently being further developed. New developments aim at making the software system available for analysis of data from similar instruments, including the Potsdam Radio Observatory (Germany). In this context, some software parts will be replaced by generic procedures already available in the Solarsoft libraries.

The Solarsoft system

The Solarsoft system (Freeland, 1999) is a set of integrated software libraries, databases, and system utilities, which provide a common programming and data analysis environment for Solar Physics. The Solarsoft system is built from HESSI, Yohkoh, SOHO, SDAC and Astronomy libraries, and draws upon contributions from many members of those projects. The Solarsoft environment provides a consistent look and feel at widely distributed institutions. It facilitates data exchange and stimulates coordinated analysis. Commonalties and overlap in solar data and analysis goals are exploited to permit application of fundamental utilities to the data from many different solar instruments. The use of common libraries, utilities, techniques and interfaces minimizes the learning curve for investigators who are analyzing new solar data sets, correlating results from multiple experiments or performing research away from their home institution.

The Solarsoft system provides a large reusable library. It is built on tracing back to SMM, through Yohkoh and SOHO, TRACE, and incorporates SXI, HESSI, and other solar observatories. Many common “solar physics” tasks are available. Solarsoft capabilities include time series analysis, time conversions, date/time manipulation (millennium safe); spectral fitting; image and image cube (movies) display; solar limb fitting; grid overlay; coordinate transformations; html conversion, file conversions, form handling, movie making, etc.

The Solarsoft system is largely hardware system and site independent. Installation and setup utilities simplify portability to different operating systems. Fundamental shared software is written in hardware-independent and site-configuration independent form. The Solarsoft setup utilities support local configuration files to get sites up and running.

The Solarsoft system provides data analysis software to synoptic data sites and primary data sites. Access to synoptic data sites from ground-based and spacecraft-based solar observatories are provided in consistent formats to facilitate coordinated analysis.

Reusable data analysis software

The concept of standardization of data-analysis software, such as implemented in the Solarsoft system, can be extended one step further. Software should be *designed for reuse*. Consider again the “theoretical” principle of data analysis: it is “just” a sequence of selections and transformations. Such a sequence can be implemented as a single pattern, and repeatedly used for each individual transformation. In other words, the data analysis process can be divided into an administrating part, called *framework*, controlling the sequence of selections and transformations, and an algorithmic part, implementing the specific transformations.

Separating the data analysis system into a framework part and an algorithmic part has the following advantages:

1. The framework is written only once, and reused for any data products. Most of the reusable part of the software is encapsulated and thus invisible for the developers of specific algorithms. They can focus on the algorithms and integrate them into a system with minimal time investment.
2. Full systems can be designed using a single framework. Each data transformation/selection is associated with an instance of the same framework, and the whole system is a sequence of these individual transformations.
3. Each transformation is standardized. It uses the same interface (that is, the set of functions used to access data) for each data product.

Framework objects

A framework can be implemented using object-oriented syntactic constructs. The framework is implemented as an *abstract class* including operations common to any data type. The class is called “abstract” because it does not contain any information about the actual data types, control parameters or algorithms involved in a specific transformation. The association of the latter with the framework is done in a *concrete class*. The concrete class inherits the framework, that is, it uses all the functionality provided by the framework. It defines data types, control parameters and algorithms associated with the inherited framework. In other words, the framework provides the functions to work with the data, and the concrete class provides the definition of data types and algorithms for a given data transformation

Framework objects have been developed and tested in the context of the HESSI data analysis software (hessi.ssl.berkeley.edu/software). They allow users of this system to access any HESSI data product in a straightforward way, even though complex transformation operations (e.g., image processing functions) are involved.

One of the challenges that frameworks face is how to deal with legacy software. Software already written should not be necessarily discarded nor rewritten. Including legacy software within newer, object-oriented software methods involve developing *wrappers* that encapsulate older data analysis systems into newer objects.

Data combination

Data combination tools must be designed to perform equivalent analysis tasks on data sets from different origins. Obviously, these tools can be based on a standard interface such as that provided by a framework object: the standard interface provides the same data access functions for any data set. This allows combining data in a simple, straightforward way, allowing users to focus on the analysis instead of having to focus on the combination itself.

Mapping software

The mapping software presented here has been developed for analyzing data from several solar instruments. This software provides methods for overlay, alignment, scaling, rotation, etc. of related data sets from the SOHO synoptic data archive (see above). In addition to data combination tasks, the mapping software also provides a common interface to access and analyze data sets from the synoptic archive.

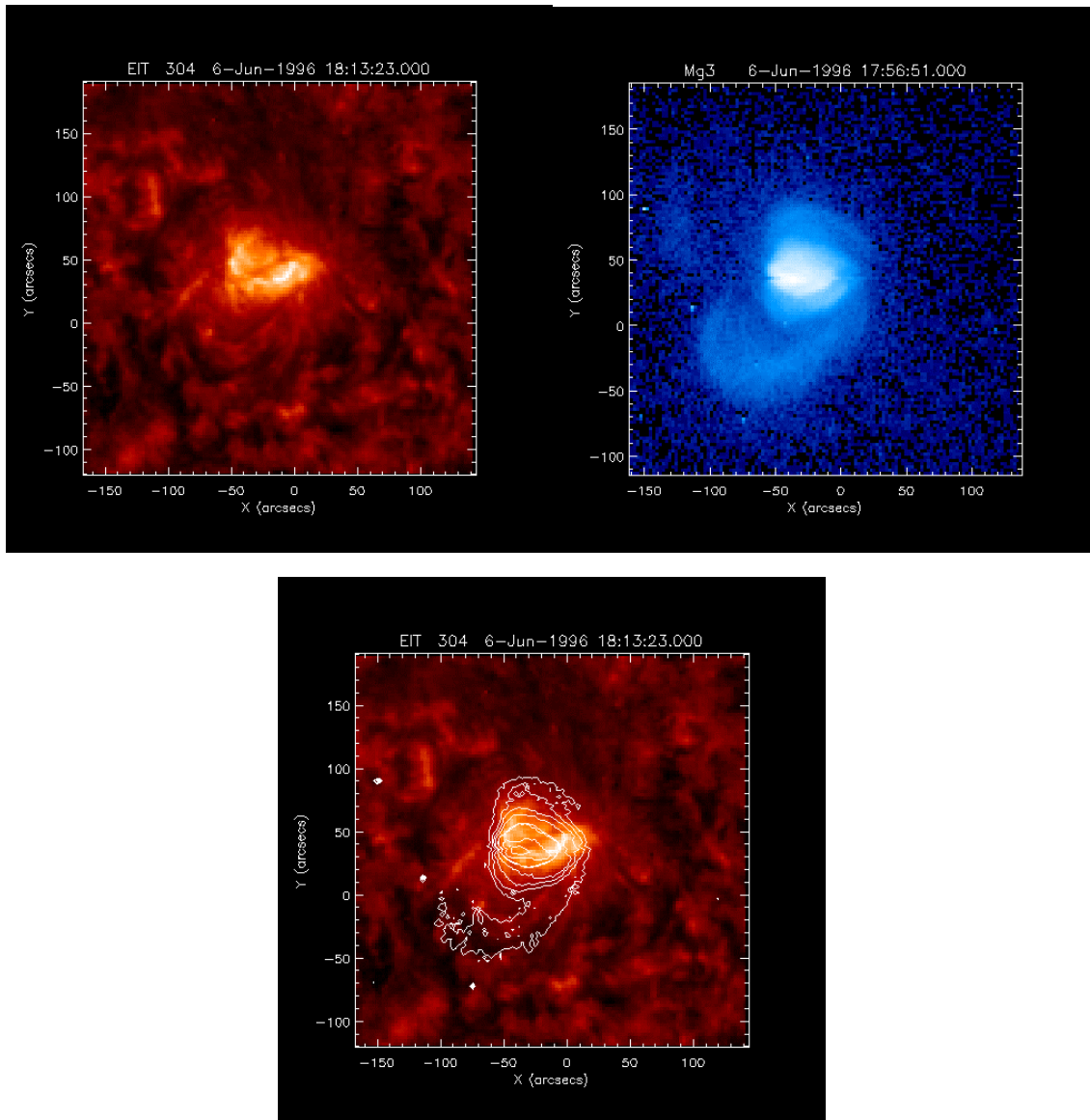


Figure 2: Example of the mapping software. An image of SOHO’s Extreme ultraviolet Imaging Telescope (EIT, upper left) is plotted next to an image of Yohkoh’s Soft X-Ray telescope (SXT, upper right). The software standardizes the pixel size and the image center such that the information from both instruments can be compared. Furthermore, the SXT image is overlaid on the EIT image with contour plots for direct comparison (bottom).

One drawback of the mapping software is that it does not allow accessing and manipulating raw data. It relies on data that have been integrated into the synoptic data site. Merging mapping software and framework objects can eliminate this drawback. Together, they can provide a way to access the data analysis software associated with a specific instrument, while keeping the ability of using the combination operations.

Visualization software

Once combined, data must also be visualized. Visualizing more than two images at the same time is challenging. Obviously, overlaying images with contour plots is limited to a small number of data sets. It does not provide a general solution. To visualize more than two images, a way to display as many images as required without overlapping has been developed. This method stacks scaled images on top of each other, by rotating them by about 70 degrees around the horizontal axis of the display area, thus introducing perspective in the representation.

Solar observations are particularly well adapted for this stacking method. We can consider a set of observations in a pseudo 3-dimensional representation where one of the dimensions corresponds roughly to altitude in the solar atmosphere. Scaled images taken at different wavelengths are shown corresponding to their most probable origin in the solar atmosphere, as shown in Figure 3.

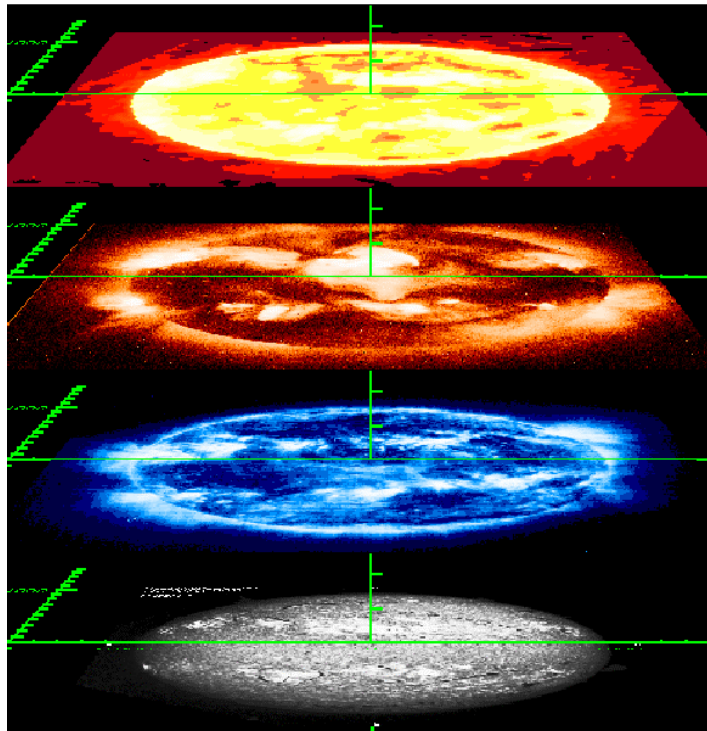


Figure 3: Four solar images are displayed using the stack representation. From bottom to top: H-alpha emission (Meudon observatory), extreme ultraviolet (SOHO/EIT), soft X-ray (Yohkoh/SXT), and radio emission, (Nobeyama radio observatory, Japan)

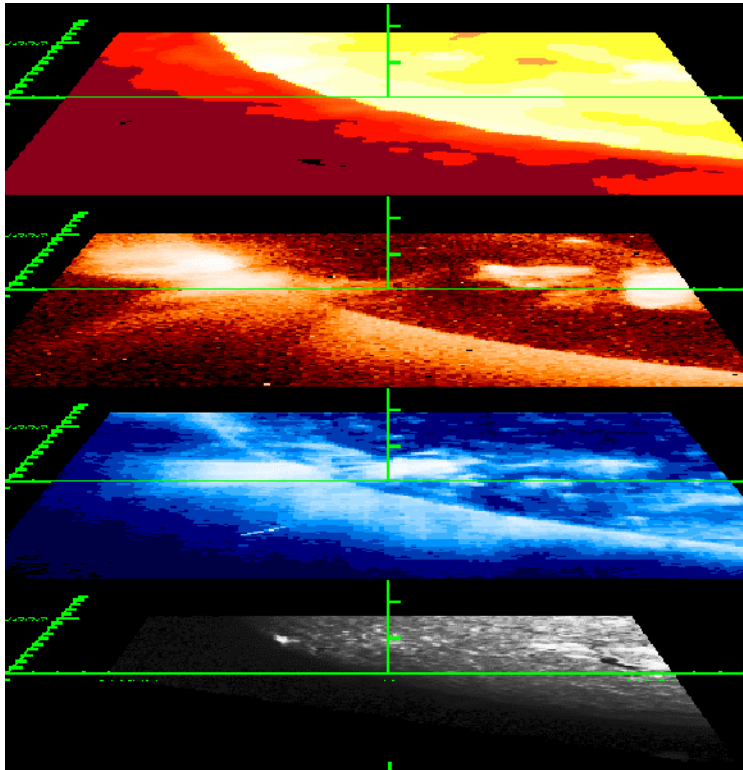


Figure 4: A zooming operation applied on all four data sets (objects) simultaneously. The images are the same as in Fig.3

The visualization software includes a set of functions that applied on all objects in the stack. For instance, regions can be selected, projected, zoomed and rotated. As a result, the representation allows seeing the relation between regions in different images, while working with as many images as wanted.

Conclusions

Archive sites today allow the access of tremendous amounts of data. But this quasi-universal *data access* does not necessarily imply efficient and effective *information access*. On the contrary. The steadily increasing amounts of archived data may bring more confusion than added value. The software to manage archived data (with some notable exceptions) often does not rely sufficiently on already developed systems. In other words, data retrieval and analysis systems for data sets from multiple instruments could return significantly more scientific results if they would rely on globally developed software. This does not apply only to solar physics, but also to scientific data analysis software in general.

The latest developments have started to “globalize” software. Developers in this field have significantly increased their efficiency by coordinating their efforts. But they need to spend even more effort in writing generic applications instead of specialized systems. By coordinating software development, they will be able to design portable, generic applications that will not only reduce the complexity of further software systems, but also contribute to robust applications. This will significantly facilitate the access to large quantities of data.

Data retrieval and data analysis systems should be considered as two pieces of a single system. This unique system should have as a goal the extraction of information from large heterogeneous data sets. It should abstract the origin of data, as well as the technical characteristics of individual instruments. Even though this seems ambitious, it should be kept in mind while designing software.

Most of the technology to develop this global scientific data handling system is already available or in development. However, the gap between available technical possibilities and their application in scientific systems is still quite wide. Projects trying to reduce this gap should be supported actively. In this way, computers will really support scientific research, and eventually build the virtual solar observatory of tomorrow.

Acknowledgements

Parts of the work presented in this article have been funded by the Swiss National Science foundation. The authors thank Brian Dennis, Säm Krucker, Christoph von Praun, Richard Schwartz, Kim Tolbert for valuable discussions, and David Ardila and Fionn Murtagh for proofreading the article.

References

- Benz, A.O., Güdel, M., Isliker, H., Miskowicz, S., and Stehling, W., 1991, A Broadband Spectrometer for Decimetric and Microwave Radio Bursts: First Results, *Solar Physics* **133**, 385-393.
- Canfield, R.C., 1999, Max Millennium Program in 1999/2000, *Bulletin of the American Astronomical Society* **31**, solar.physics.montana.edu/max_millennium
- Csillaghy, A., 1997: Information Extraction by Local Density Analysis, PhD Thesis, Shaker Verlag (www.shaker.de).
- Csillaghy, A., 1998: The Data Analysis Software Ragview, www.astro.phys.ethz.ch/rapp/software/ragview-manual.ps
- Csillaghy, A. and Benz, A.O., 1999, Interactive Image Retrieval in Large Astronomical Archives: the ASPECT System, *Solar Physics* **188**, 203-216.
- Freeland, S.L., 1999, The Solarsoft System, www.lmsal.com/solarsoft
- Messmer, P., Benz, A.O., and Monstein, C., 1999, PHOENIX-2: A New Broadband Spectrometer for Decimetric and Microwave Radio Bursts: First Results, *Solar Physics* **187**, 335-345.
- Perrenoud, M., 1982, The Computer-Controlled Solar Radio Spectrometer IKARUS, *Solar Physics* **81**, 197-203.
- Wells, D.C., Greisen, E.W., and Harten, R.H., 1981, FITS: A Flexible Image Transport System, *Astronomy and Astrophysics* **44**, 363-370.
- Zarro, D.M., Canfield, R.C., and Csillaghy, A., 1999, The HESSI Coordinated Data Analysis Flare Archive, *Bulletin of the American Astronomical Society* **31**.