*SF2A 2006*
*D. Barret, F. Casoli, T. Contini, G. Lagache, A. Lecavelier, and L. Pagani (eds)*

# DATA COLLECTION AND MINING IN SOLAR PHYSICS

J. Aboudarham[1]

**Abstract.** In solar physics, data are scattered in several places which can be databases as well as archives. Observations at various wavelengths give informations on the vertical behavior of Solar atmosphere. But the organization of observing campaigns with various intruments (ground and/or space-based) is very complicated in order to ensure simultaneous observations. Another way to handle this difficulty is to search in various archives observations that fit together. New opportunities to study the Sun are offered by recent technics of automatic feature recognition, which allow data mining that can be very rich when applied to several archives. All those questions can be addressed using the concept of virtual observatory, that is already developed in Solar Physics.

## 1 Introduction

Solar observations are not subject to the same constraints as night time observations. So several solar observatories continue to provide observations even if they are located close to towns. This explains why solar data are so scattered all around the world.

The complementarity between ground-based (generally in visible or radio wavelength) and space-based (UV and X mostly) observations is very important as they give informations on various altitudes in the solar atmosphere. It is now quite common - but not so easy to put into practice - to have coordinated observations between several instruments in order to understand the vertical behaviour of the atmosphere.

A new emerging subject in solar physics is the automatic feature recognition which allows to automatically detect solar structures on - generally - full Sun images.

All this makes that data collection and mining in Solar physics is at a turning point. One must take that into account when collecting data from either space- or ground-based instruments.

## 2 Solar data

The status of solar data depends on their origin. Space data are generally clustered by satellite/probe and, in the case of the biggest data centers (such as MEDOC or NASA) several instruments can share the same database.

The case of ground-based observation is much more complicated. In most places, each observatory has - at best - its own database, with the noteworthy exception of France, where all solar observations are archived in the BASS 2000 database (http://bass2000.obspm.fr).

In many observatories, scientists return to their laboratory with their observations, which are then lost for the solar community.

The file format used for solar observations is generally FITS format for all space-based data and several ground-based ones. But there are still instruments that deliver files in a proprietary format.

Another characteristic of some solar data is that the files may be quite big and contain in some cases several hundred, indeed thousands of individual images. This happens generally for raster scans or long duration observations of a specific region on the Sun.

All these elements have to be taken into account when one wants to relate solar data scattered in several places.

[1] LESIA, CNRS and Observatoire de Paris, 5 place Janssen, F–92195 Meudon cedex, France

## 3 Solar physics needs

### *3.1 Classic needs*

As each observing wavelength corresponds to specific physical characteristics in the solar atmosphere, the understanding of its vertical behavior implies the use of simultaneous observations at various wavelengths. For instance, one can roughly say that visible wavelength give informations on the lower atmosphere (photosphere and chromosphere), EUV and X-rays on transition regions and low corona, and radio waves on the extended corona.

Moreover, the fact that many solar phenomena are transient forces to have a very good temporal coordination of observations. Then several scientists prepare coordinated observations using both ground- and space-based instruments. These programs encounter two main kinds of problems:

1. Requests for observing time have to be made long time before the concerned period, and have to be approved for the same moment by all the instruments' scientific committees

2. When ground-based instruments are involved, no one can guarantee that the weather will permit observations. Constraints for solar observations are very different from night sky observations: not only has the sky to be clear, but also one must avoid turbulence which occurs often when the Sun shines!

### *3.2 New needs*

As one of the main research programs concerns the understanding of the solar magnetic field behavior, scientists often try to make observations of solar events perturbing the magnetic field. As mentioned earlier, the local and transient aspects of solar phenomena prevents from being sure to observe them. But to miss an event does not mean that nobody has ever observed it!

There are many systematic observations of the Sun all around the world (and even from space). They hold many useful informations if one looks at them day after day, from one place to the other. For instance, filament disappearance can be detected that way, as well as several other phenomena of interest. But it is a heavy task to scan all systematic observations, that cannot be completed totally with the new generation of high resolution observations.

Solar-Terrestrial relationship is in need of new statistical tools that can reveal the way phenomena originating in various locations and at various scale can be related together.

Last but not least, the new generation of solar instrument (for instance the "Solar Dynamics Observatory" probe or the "Frequency Agile Solar Radiotelescope") will deliver a tremendous amount of data, more than one Terabyte a day! No one can check manually such an amount of data. And there is now no evidence for the possibility to store all the data. This means that an automatic selection of data will have to be made.

## 4 Issues and answers

### *4.1 The prototypes*

In the US, the Virtual Solar Observatory, VSO, (http://www.virtualsolar.org/, see Gurman et al. 2005) provides an entry point for an easy access to 13 archives, 12 in the US, together with BASS 2000 at Meudon. It is supported by NASA.

In Europe, a very strong effort has been made, taking benefit of a three-years support of the European Union, leading to the realization of EGSO, European Grid of Solar Observations (http://www.egso.org, see Bentley et al. 2004). This is a grid-based virtual observatory. It proposes access to 18 archives, in US, south America and Europe, a Solar Event Catalogue (SEC) merging informations from 17 catalogues of solar events (CME, flares, proton events ...), and a Solar Feature Catalogue (SFC) providing a database of filaments, chromospheric active regions and sunspots based on an important effort concerning automatic feature recognition (see Zharkova et al. 2003 and Fuller et al. 2005 for instance on that topic). Note that some of the data are queried using the US VSO as a data provider.

One of the most original features of EGSO is the possibility to drive a query using solar events. This means that it is possible to look for observations that correspond to listed solar events: for instance, one can look for the important proton events during a period of time and get a list of observations during the selected proton

event. A Java data browser, working with several file formats, including FITS, has also been developed, allowing to visualize data before downloading them.

### *4.2 Attempts to answer*

Several issues concerning solar observations have been addressed at the beginning of this paper. The prototype virtual observatories (VOs), VSO and EGSO, provide the beginnings of an answer to that.

**3-D informations on the solar atmosphere** : this point can be solved by VOs which allow wide selection criteria. It is easy to get a complete set of observations that occurred exactly at the same moment.

**Non-uniform data format** : VOs, in order to work, need to hold a dictionary that translates the"language" of an archive in some common language which can be applied by the user to access several archives. As VOs become more used, this will develop a standard that will be applied to future instruments.

**Search for events** : a solar event catalogue merging all the available catalogues allows to build sophisticated queries that were very difficult to build before. A solar feature catalogue has also been developed (see Fig.1 for a use of it). But it has to be extended to new solar features, and include a time tracking of solar structures. This will open a new chapter in the study of the Sun by providing the necessary data for advanced studies of the behavior of those structures.

**Solar-Terrestrial relationship** : new means to tackle this issue will be the result of the development of solar feature catalogues. Then extended statistical studies of the behavior of solar structures and/events related to activity will be possible, and even to cross-correlate them in long periods of time.

**Big amount of data** : once again, the developments in solar feature recognition technics will provide the tools necessary for automatic selection of data of interest; both at source, for the choice of data to archive, and on the consumer side, to detect which data correspond to the interest of the user.

## 5 Conclusion

The complexity of the study of the Sun and its influence in the solar system, as well as the huge amount of data that the new generation of solar instruments will provide, imposes to develop new tools that help to query and recover appropriate data. The natural answer to this is the creation of a virtual observatory of solar interest together with tools for information extraction to provide non trivial qualities of service.

A first step has been cleared with the creation of the European EGSO and US VSO. But it has to be carried on, especially concerning automatic feature detection and mass data processing.

## References

Bentley, R. D., Csillaghy, A., Scholl, I., et al. 2004, 35th COSPAR Scientific Assembly, held 18 - 25 July 2004, in Paris, France., p.3935

Gurman J.B., Dimitoglou G., Hourcl J., et al. 2005, The Virtual Solar Observatory: Still a Small Box, AGU

Zharkova, V. V., Ipson, S. S., Zharkov, S. I., et al. 2003, Solar Physics, 214, 89

Fuller, N., Aboudarham, J., Bentley, R. D. 2005, Solar Physics, 227, 61

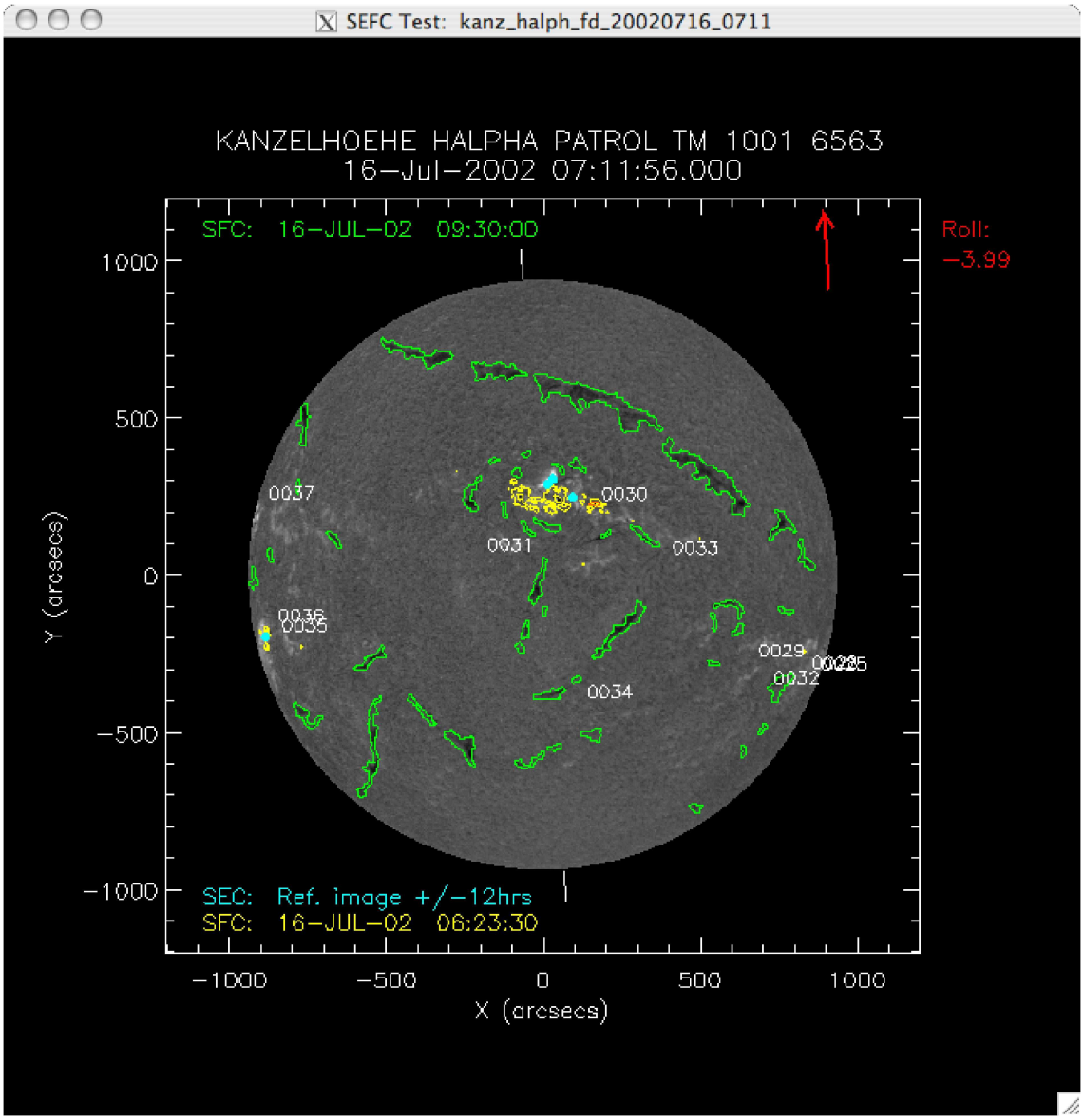


**Fig. 1.** On this image, solar structures automatically detected on Meudon observations have been superimposed on a Kanzelhhe image, as well as informations extracted from the EGSO Solar Event Catalogue. The next step of a virtual observatory would be to be able to click on a structure and retrieve a list of its observations.