Mem. S.A.It. Vol. 74, 823

Memorie della

EGSO - The European Grid of Solar Observations

K. Reardon[1], E. Antonucci[2], S. Giordano[2], G. Severino[3], M. Messerotti[4], and the EGSO Team

[1] INAF/Osservatorio Astrofisico di Arcetri
[2] INAF/Osservatorio Astronomico di Torino
[3] INAF/Osservatorio Astronomico di Capodimonte
[4] INAF/Osservatorio Astronomico di Trieste

Abstract. The European Grid of Solar Observations (EGSO) project aims to produce the framework for a coordinated community-wide resource for obtaining and reducing solar observations. The EGSO will be capable of sharing resources coming from all types of providers, while ensuring scalability, security, and compatibility among all datasets. The user will be provided with a customizable search interface from which to simultaneously browse or data mine a range of solar and heliospheric data archives. In essence, the EGSO will create the fabric of a virtual solar observatory.

Key words. Virtual Observatory – Data Bases – Data Analysis

1. Introduction

Solar Physics is a data rich discipline, generously endowed with a wealth of photons by a prodigious patron. The role of the European Grid of Solar Observations (EGSO) is to provide the infrastructure, based on concepts such as grid sharing of resources and peer-to-peer networking, that will allow researchers to better exploit the whole range of available information. The system will provide the means to map solar data and metadata onto a common framework to greatly simplify the comparison of the information in multiple heterogeneous repositories. New modes of interaction with the data will be enabled to reduce the time spent locating and retrieving data of interest and allow more time for the analysis of those data.

EGSO is funded by the European Commission as part of the Information Society Technologies Program of the Framework Programme 5. The project involves partners in Italy, UK, France, Switzerland and the US (Bentley, 2002). It is a parallel effort to projects such as the Virtual Solar Observatory (VSO) in the United States (Gurman, 2002), and the Astronomical Virtual Observatory (AVO) led by ESO. The italian solar physics community participates through the Instituto Nazionale di Astrofisica, led by the Osservatorio Astronomico di Torino,

Send offprint requests to: K. Reardon, kreardon@arcetri.astro.it

with the explicit inclusion of other italian observatories (Messerotti, 2003).

2. Motivations

The realization of a distributed system for resource sharing within the solar physics community is driven by several factors. We describe here some of the issues that must be dealt with to provide a long-term solution to these problems.

Distributed Data Sources: It is inherent to the acquisition of solar data that the information obtained remains distributed at multiple locations spread throughout the world. Solar instruments installed at different locations are almost always unique implementations of a particular concept, for which a certain amount of detailed knowledge may be required to interpret the acquired data. Instruments operated with local resources must generally retain the local control of the data obtained for political and practical reasons. At the same time, there are a range of reasons for which the usage of a number of data sources is advantageous in solar physics. For example, combining observations from instruments spaced around the globe is especially useful in the study of phenomena that vary on timescales comparable to the period of the earth's rotation. Combining multiple data sources is often necessary to study emissions from different regimes in the solar atmosphere, and construct a three-dimensional picture of phenomena that span a range of temperatures and densities.

Large Data Volumes: The recent increase of spectral and spatial resolution, without loss of overall coverage, is greatly expanding the volumes of data that must be recorded. Even missions or instruments that operate at relatively low data rates may, operating over decades, obtain scientifically useful datasets that grow to prodigious sizes over time. Current heliosesimological datasets already contain tens of terabytes. New instruments, such as The Solar Dynamic Observer or the Fast-Agile Solar Radiotelescope (FASR) are designed to generate several terabytes of data per day continuously over many years. The data volumes, pushing into the petabyte regime, are comparable to those envisioned for the Large Hadron Collider project at CERN.

Data Mining: The increasing wealth of data makes it ever harder to navigate the vast sea of available information. Locating the combination of data from multiple instruments most suitable for addressing a given problem is increasingly problematic. Due to the effort involved one rarely exploits the complete range of data available, relying on a few examples of a given phenomena, rather than compiling a more statistically significant sample. Furthermore, there may be additional phenomena or correlations of which we may only become aware through the exploration of the data using large number of measurements. This type of study is a form of data mining, which relies on the ability to manipulate and explore large quantities of data, taking slices over multiple dimensions or parameters.

3. Archive Federation and the EGSO Architecture

The motivations outlined above lead to the need to federate the existing data archives into a virtual whole. Such a federation would not modify or replace individual archives, but would provide a layer of abstraction above them in such a way that all existing archives would be presented to the scientists in a unified format.

A federation of this kind requires the knowledge of the way in which the contents of each metadata catalog or data file are described by different data archives. It also requires the definition of a common framework onto which the contents of all the included catalogs can be mapped. Since each archive remains separate and distributed, there must be the means to convert a general request from the researcher into the single queries to be transmitted to the individual archives in order to construct the answer to the user's request. In this way, the complexity and variety of data resources can be handled by the system itself without placing the burden on the user.

The architecture designed for EGSO envisions three separate *roles* that will handle these various tasks. These roles are: *Consumer*, to handle user-related interactions and processes; *Provider*, to provide flexible access to the offered data, metadata, and services; *Broker*, to provide the means to locate, query, and translate the providers in the name of the users. These three roles define the boundaries between the various software components of the system, but need not necessarily constrain the actions taken by individual participants in EGSO. In addition, it is not required that the Consumer or Provider software components reside locally on the computers of the actual clients or data sources for the EGSO system.

3.1. Provider

Viewed as a supply chain, the provider is the source of all that is handled by EGSO to satisfy user requests. The Provider is the crucial interface between the existing data archives and the rest of the EGSO system. A "provider" may offer one or more of the following types of resources: Metadata, descriptions of observations obtained, as well as other extracted information about the solar conditions or behavior; Data, the actual data files, in raw or reduced form, that were obtained from observations; or Services, being any capabilities, such as pre-defined software routines, processing capabilities, or knowledge repositories that are offered in a well-defined fashion to the EGSO users. These three different elements are handled by three different subsystems within the Provider role. The goal of these different subsystems is to provide a uniform method of access to all the archives that participate in EGSO.

3.2. Broker

The Broker inserts itself as an intermediary between the provider and the consumer in order to obscure the multiplicity and heterogeneity of the providers from the end users. By having the broker handle the dirty work of resource location and homogenization, the capabilities required of the Consumer can be made simpler and easier to implement in the variety of conditions found at the end users. There will be multiple Brokers implemented, above all to allow for redundancy or efficient access to a distributed user base, but also possibly to allow for specialized brokers that offer information on specific resources or provide certain types of access (e.g. for data requiring authorization).

The Broker will maintain a list of available providers, a resource registry, based on information submitted by the providers themselves, as well as information shared from other Brokers. The information will include the details to permit mapping of the provider's metadata catalog contents onto a common data model in order to allow the comparison of catalog contents from multiple providers. The Broker will also maintain a catalog, at course granularity, containing a summary of all the observation catalogs that have been incorporated into EGSO. This meta-catalog will allow for a rapid identification of those data resources that are likely to have information relevant to the user's query.

3.3. Consumer

The Consumer provides the means for a user of the system to actually visualize, select, and manipulate the metadata and data obtained from multiple resources. Users accessing EGSO will make requests to the system for accessing or utilizing the various resources (Metadata, Data, Services) made available by the providers. The Consumer software will be responsible for accepting those requests, expressing them in a manner appropriate for the system as a whole, and then passing the requests on to the Broker. The Broker will then pass the result generated according to the user's request back to the Consumer, which will then present the results to the user in their desired fashion.

The Consumer will be able to store the results provided by the Broker, both in memory or, for longer term persistence, in a disk cache managed by the Consumer. The Consumer components will then be able to operate on these cached information, for example to increase system response during the user's successive refinement of their initial search parameters, or to allow for some system functionality even in the absence of network connectivity. Since all the information returned by the Broker will be presented according to a single common representation, the Consumer will simply need the ability to work with this one data model. The Consumer will also offer the possibility to use or define *workflows*, sequences of operations to be performed upon appropriate datasets of interest.

4. Scientific Application

This underlying architecture will allow the construction of a complete system granting solar physicists new powers in the identification of events of interest. Once a scientist has selected a research topic to be addressed, the next step in the process is to identify what existing data might best offer insight into the problem at hand. EGSO will provide the ability to search multiple catalogs rapidly with single query, reducing the time spent on mundane search tasks. Correlative searches across multiple catalogs (e.g. find those occasions when two instruments took images within one minute of each other), currently all but impossible except in some special cases, will be broadly available. The ability to run pre-defined, or even user-provided, software on a given set of data, possibly remotely at the archive site itself where the entire dataset is readily available, will facilitate the selection of data based on the information contained in the data themselves.

Finally, one component of the EGSO project is to apply feature recognition techniques to solar data with the goal of creating new catalogs of features in the solar atmosphere(e.g. filaments, CME's). These new catalogs will be an immediately useful product for the solar physics community, as will the software produced that can be applied to other data or features. At a deeper level, this aspect of the project is a testing ground for the power that should ultimately be made available to the community. This would include the capability to apply a user-created workflow or program to one or more datasets, possibly at the archive itself or on a third-party computational resources, in order to derive a new catalog of features on the sun, a catalog that can then be fed back into the system to find data obtained during the duration of the feature. This power to define, algorithmically, a new solar structure, find occurrences of that feature, and obtain the data pertinent to the identified events, possibly all in the space of a few hours, will be of fundamental importance in truly exploiting the vast quantity of information contained in the solar data.

References

Bentley, R.D., EGSO Consortium: 2002, *EGSO - the European Grid of Solar Observations*, in 10th European Solar Physics Meeting, A. Wilson. ed., ESA SP-506, Vol. 2, p. 923 - 926.

Gurman, J.D., 2002, *Toward a Virtual Solar Observatory* in Proceedings of the SOHO 11 Symposium , A. Wilson, ed., ESA SP-508, p. 525 - 528.

Messerotti, M., et al., 2003, *The Italian Solar Data Archives: National and European Perspectives*, Mem. SAIt., in press.