EGSO - The Fabric of a Virtual Archive

R.D. Bentley, K. Reardon
and the EGSO Consortium

For more information, contact us on:
egso@mssl.ucl.ac.uk

The EGSO Web site can be found on URL:
http://www.mssl.ucl.ac.uk/grid/egso

What is EGSO?

EGSO, the **"European Grid of Solar Observations"**, is a Grid test-bed that will change the way users analyse solar data. The project is funded under the **Information Society Technologies** (**IST**) thematic programme of the European Commission's Fifth Framework Programme (**FP5**). It was approved in the autumn of 2001, following the submission of a proposal earlier that year. **EGSO** started in March 2002 and will last for 3 years.

One of the **major hurdles in the analysis of solar data** is finding what data are available and retrieving those that are needed. Planned space- and ground-based instruments will produce huge volumes of data and even taking into account the continuous technical advances, **it is clear that a new approach is needed to the way we use these data**.

EGSO will federate solar data archives across Europe and beyond to produce a virtual data set. It will provide a friendly user interface for the tools needed to select, process and retrieve the distributed and heterogeneous solar data. **EGSO** will provide mechanisms to produce Unified Observing Catalogues for space and ground-based observations, and will also provide the tools to create Solar Feature Catalogues that will create a new way to select solar data based on features, phenomena and events. In essence, **EGSO will provide the fabric of a virtual observatory**.

The Partners in EGSO

University College London (UK) **Co-ordinating Group**

- Dept. of Space and Climate Physics (MSSL)
  Contact Person: Bob Bentley **Project Co-ordinator**
- Dept. of Computer Science
  Contact Person: Anthony Finkelstein

Rutherford Appleton Laboratory (Didcot, UK)

- Dept. Space Science and Technology
  Contact person: Dave Pike

University of Bradford (UK)

- Dept. of Cybernetics
  Contact person: Valentina Zharkova

Osservatorio Astronomico di Torino (Turin, Italy)
Includes the Observatories of Naples and Trieste
Contact person: Ester Antonucci

Politecnico di Torino (Turin, Italy)

- Dept. di Automatica e Informatica
  Contact person: Luigi Ciminiera

Observatoire de Paris-Meudon (Meudon, France)
Contact person: Jean Aboudarham

Institut d'Astrophsique Spatiale (IAS) (Orsay, France)
Jointly funded by CNRS & Universite Paris Sud
Contact person: Isabelle Scholl

University of Applied Sciences (Brugg-Windisch, Switzerland)

- Dept. of Computer Science
  Contact person: Andre Csillaghy

Solar Data Analysis Center, NASA-GSFC (Washington DC, US)
Contact person: Joe Gurman

National Solar Observatory (Tuscon AZ, US)
Contact person: Frank Hill

Note: UAS, SDAC and NSO do not receive funds from the EC

Associated Partners

Astrium plc (an Anglo/French/German company)
Lockheed-Martin (Palo Alto CA, US)
ESA SoHO Project (GSFC, Washington, US)

The Data

Solar data are scattered around the world. In order to solve many of the problems in solar physics, scientists need to be able to access as wide a range of wavelengths as possible. These are used to construct a (5D) picture of how the density and temperature of plasma held in structures in the solar atmosphere changes with time. For technical reasons, optical and radio observations are normally made from the ground, while UV, EUV, X-ray and γ-rays are observed from space.

- **Space-based observations** are often held in data centres that contain multiple data sets that are readily accessible from the Internet. This capability has developed over several years as a consequence of the international collaborations needed to build the instruments on numerous missions.

- **Ground-based observations** are made by institutions that are scattered across many time zones and usually act independently. Although a few observatories have developed archives (e.g. Obs. Paris-Meudon and NSO), many are resource limited and not used to sharing data.

The **data are very heterogeneous**. Also, existing cataloguing of the observations is very inconsistent and is often dependent on ancillary data. Many of the data are in an un-calibrated form and are calibrated on the fly as they are extracted – **SolarSoft** is used extensively by the community for this. Although most is in the public domain, **some data are proprietary**.

New instruments currently in the planning stages will produce **volumes of data** significantly greater than at present. EGSO will provide access that will reduce the need to distribute this. To ensure optimal response times, a limited number of copies will need to be placed in strategic locations, but that is all.

The Users

The users of solar observations are diverse and are scattered around the world. Even in the solar physics community, while some users are well known to us, others are totally unknown. Other potential users of solar data (e.g. space weather, climate physics and astrophysics) come from communities that are generally unknown to us. Control of user access is one of the problems confronted by **EGSO**.

User authentication and **authorization** are therefore serious issues, particularly in relation to access to and protection of the data:

- Data providers need to be sure that their data are secure and that access rights will be respected
- Users want to know that the intellectual property of their work – the results of database searches and data products – will be protected. Also, often they want to be able to share data with collaborators, but deny access to others.

The facilities available to the users differ greatly. An objective of **EGSO** is to enhance access to an extended set of data while reducing the need for elaborate local facilities. Processing the data at source is key for this.

Overview of the Application

Identify suitable observations

Many solar observations are serendipitous – instruments observe the Sun and interesting events and phenomena are then analyzed. The data are heterogeneous and catalogues are key in tying them together. Current catalogues differ in quality, contents, dependencies and format. **EGSO** will produce a new set of **Unified Observing Catalogues** that are self-describing (e.g. using XML), quantized into fragments (by instrument and time), and with all dependencies on ancillary data and proprietary software removed. **Solar Feature Catalogues** will provide a new entry point into solar data based on features, events and phenomena rather than just date, time, wavelength and pointing.

Locate the data

Once the desired set of observations has been identified, they need to be located – *this is a different problem to knowing their contents*. The **data are scattered**, with differing means of access. There may be multiple copies of the data, each with different currencies and degrees of integrity – which is most easily accessed will change with time. Also, while most data are public, others are proprietary.

Process the data

Processing involves the extraction and calibration of the data. **SolarSoft** provides set of common analysis tools written mainly in IDL – calibration information is distributed within this environment. Often the user only needs a subset of each data set, and given the increasing volumes and complexity of the data, we require a sea change in approach to the way data are handled. An objective of **EGSO** is to reduce the volume that is returned to the user by extracting at source. For complex analysis tasks, **the user will be able to upload their code to data sources**.

Return results to the User

All data will be returned in files that are self-describing, possibly an XML version of the FITS file standard. This will facilitate the use of data products by other user communities.

Compare results from different instruments

Tools to compare observations already exist within **SolarSoft**. The main difference will be that the data will already by extracted and calibrated. This will allow the user to compare more diverse data without the necessity of a complex installation of **SolarSoft**.

EGSO Work Packages

Catalogues (WP4 & WP5)
Catalogues are the means of identifying what observations are available. **EGSO** will create a set of **Unified Observing Catalogues** (WP4) to facilitate the search for matching observations, and sets of **Solar Feature Catalogues** (SFCs; WP5) to provide a new entry point into solar data based on features, events and phenomena.

- Some SFCs will be created using image recognition software to identify features in full disk imagery at different wavelengths
- Other SFCs will be created from published lists of events, etc.

Search and visualization tools (WP3)
An interactive User Interface (UI) will be provided to facilitate the search of catalogues for suitable observations – summary and synoptic images and other ancillary data will assist the user in their choice. Since different users may wish to search on different criteria, it will be possible to modify the elements of the UI with a user preference file.

Data retrieval and processing (WP2)
The location of data is a different problem to their contents. Tools will be provided to locate the best source of requested data; extract the subset, process and return them to the user. Extraction will be mainly at source; users will be able to upload code to data sources.

System Definition and Integration (WP1)
This WP will determine the overall requirements and constraints of the project – these will be used to produce the Architectural Design of **EGSO**. The WP will integrate the components generated in the WPs 2, 3, 4 & 5 to create the **EGSO**.

Project Control and Dissemination (WP0)
Project management and Dissemination are very important parts of EGSO and are grouped in WP0. Also included in this WP is the assessment and evaluation of other WPs.

Relationship to Other Grids

The US Virtual Solar Observatory (VSO)

Following a recommendation by the NASA Senior Review in the early autumn of 2001, funds to help develop a virtual solar observatory are being made available through the SDAC at GSFC. It is hoped that work will start shortly on the **VSO** following the recent review of proposals submitted under this initiative. The **EGSO** Team includes US partners to try to ensure that Europe and the US work in tandem on creating this important project. **EGSO hopes to work closely with the VSO project**.

Notes:

1) *Lockheed-Martin is working on a similar project to VSO funded under the "Living with a Star" program. EGSO are already in discussion with the team in Palo Alto.*
2) *There will be a session at AAS/SPD meeting in June on the virtual observatory concept.*

GRIDSTART (An EC-funded IST project)

Ten Grid projects have been funded under the IST thematic programme of FP5. In order to ensure a maximum return to the IST programme, these projects have been grouped into a cluster within the Accompanying Measure project GRIDSTART – this is lead by the EPCC. EGSO is a partner in GRIDSTART, enabling to gain access to information and expertise of the other projects.

SpaceGRID (An ESA study project)

Funded by ESA General Studies, the **SpaceGRID** project aims to assess the requirements in four different ESA technical areas (Space Science, Earth Observation, system engineering and modelling), the identification of the initial ESA-wide infrastructure and the prototyping of specific components in the different technical areas. **EGSO** and **SpaceGRID** are cooperating on a no exchange of funds basis, discussing user requirements and sharing knowledge of Grid middleware, etc.

AstroGrid (A UK PPARC-funded project)

Within the UK, a Grid project to federate astrophysical, solar and STP data has begun – this (apart from the solar and STP parts) is the analogue of the EC-funded Astrophysical Virtual Observatory (AVO) and the US National Virtual Observatory (NVO). **EGSO** and AstroGrid are cooperating closely – UCL is responsible for the solar part of AstroGrid, and RAL is also heavily involved.

Relevance of EGSO to STEREO

EGSO is intended to change the way researchers access solar data:

- Under **EGSO**, it will be possible to carry out sophisticated searches for observations before the data are retrieved. Possible searches will include a wide range of space and ground-based observations.
- The user will not need to know where the data are located, the system will take care of this and will choose the data sources that are optimal for the user at the time of the request.
- Because **EGSO** will provide enhanced access to the data, the way in which data are distributed within a project will change. There will no longer be the need to make copies for each CoI institute, avoiding all the problems inherent in this. Given the volume of data that **STEREO** will produce, the capabilities of **EGSO** could be very important.

The **EGSO** Team wants make it simple to add data sets to the data Grid it will produce. An early objective is therefore to define **standards for data file headers** that will ensure the compatibility of future data sets with the **EGSO** and **STEREO** projects. The standard will be developed in discussion with interested parties, including the **STEREO**, **Solar-B** and **ILWS** projects. It will ensure that the file headers include an minimum set of standardised parameters that will simplify the production of the light-weight catalogues that are needed for the search tools **EGSO** will provide to conduct extended searches.

STEREO poses particular problems because the observatories are significantly away from the Sun-Earth line and their location changes with time as the spacecraft drift away from the Earth. Although **STEREO** will be the first major observatories of with these characteristics, similar deep-space missions are planned and it is important to ensure that **EGSO** can cope with this type of observatory.