European Grid of Solar Observations

Management of Metadata in virtual solar observatories – Experiences from EGSO

Bob Bentley (UCL) & the EGSO Team

16 December 2003
IST Workshop on
Metadata Management in Grid & P2P Systems

egso

GRIDSTART

UCL

Information Society Technologies

Outline

- Overview of EGSO
  - The problem being addressed by EGSO
- EGSO Search capability
  - Solar Event and Feature Catalogues
- Current status of EGSO

EGSO – European Grid of Solar Observations

- EGSO is a Grid test-bed related to a particular application
  - Designed to improve access to solar data for the solar physics and other communities
  - Addresses the generic problem of a distributed heterogeneous data set and a scattered user community
- Funded under the Information Society Technologies (IST) thematic priority of the EC's Fifth Framework Program (FP5)
  - Started March 2002; duration of 36 months
- Involves 11 groups in Europe and the US, led by UCL-MSSL
  - 4 in UK, 2 in France, 2 in Italy, 1 in Switzerland, 2 in US
  - Several associate partners, mainly in the US
- EGSO is interacting with many other projects
  - Working closely with equivalents in US – VSO & CoSEC
    - Joint Technical Meeting in December 2003 in San Francisco
  - Collaborated with ESA's study project SpaceGRID
  - Involved with other EC funded Grid projects through GRIDSTART

Objectives of EGSO

- **Support user community scattered around the world**
  - Current and future projects are international collaborations
  - EGSO funded by EC, but has US partners
- **Build enhanced search capability for solar data**
  - Analysis of solar data is often **event driven**
  - Search capability linked to this not currently available
  - Increasing data volumes, etc. require new methodology
- **Provide access to solar data centres and observatories around the world**
  - Data available in Europe (or US) not enough for many studies
- **Provide ability to process data at source**
  - Both pipeline and more complex processing
  - Includes ability to upload code to some providers

Need for these capabilities not unique to solar physics…

The extended EGSO family

Partners provide expertise in solar physics and IT

- **UK**
  - UCL-MSSL, UCL-CS, RAL, Univ. Bradford, Astrium
- **France**
  - IAS (Orsay), Observatoire de Paris-Meudon
- **Italy**
  - Istituto Nazionale di Astrofisico, Politecnico di Torino
    - INAf includes observatories of Turin, Florence, Naples and Trieste
- **Switzerland**
  - Univ. Applied Sciences (Aargau)
- **Netherlands**
  - ESA – Solar Group
- **US**
  - SDAC (NASA-GSFC), National Solar Observatory (VSO)
  - Stanford University, Montana State University (VSO)
  - Lockheed-Martin (CoSEC)

Current Status of EGSO

- Extensive survey of requirements in 2002
- Working architecture defined and detailed during the first half of 2003

- Release 1 of EGSO was demonstrated at IST2003 in Milan (October 2-4)
  - Demonstration of how the three roles work together
  - Working prototype of the Solar Event Catalog Server
- Release 2 of EGSO due at the end of November
  - Development of interface to Data Providers
  - More complex query supported through SEC Server
  - Greater GUI capabilities

- New version of EGSO Data Model document was released recently
  - Describes both solar and heliospheric data

Generic Solar Physics Query

- **Identify suitable observations** *(many serendipitous)*
  - As many different data sets as are available
  - Should be possible without accessing the data
- **Locate the data**
  - Data scattered, with differing means of access (some proprietary)
  - Often only need a subset of each data set
- **Process the data**
  - Involves extraction and calibration of a subset of *raw data*
  - Uses code defined by instrument teams (*SolarSoft*, C…)
- **Return results to the User**
- **Compare results from different instruments**
  - *SolarSoft* (IDL) provides a standard platform for analysis

Note the exchange in order of the 3rd and 4th bullets in this Grid expression of the problem, as compared to current practice

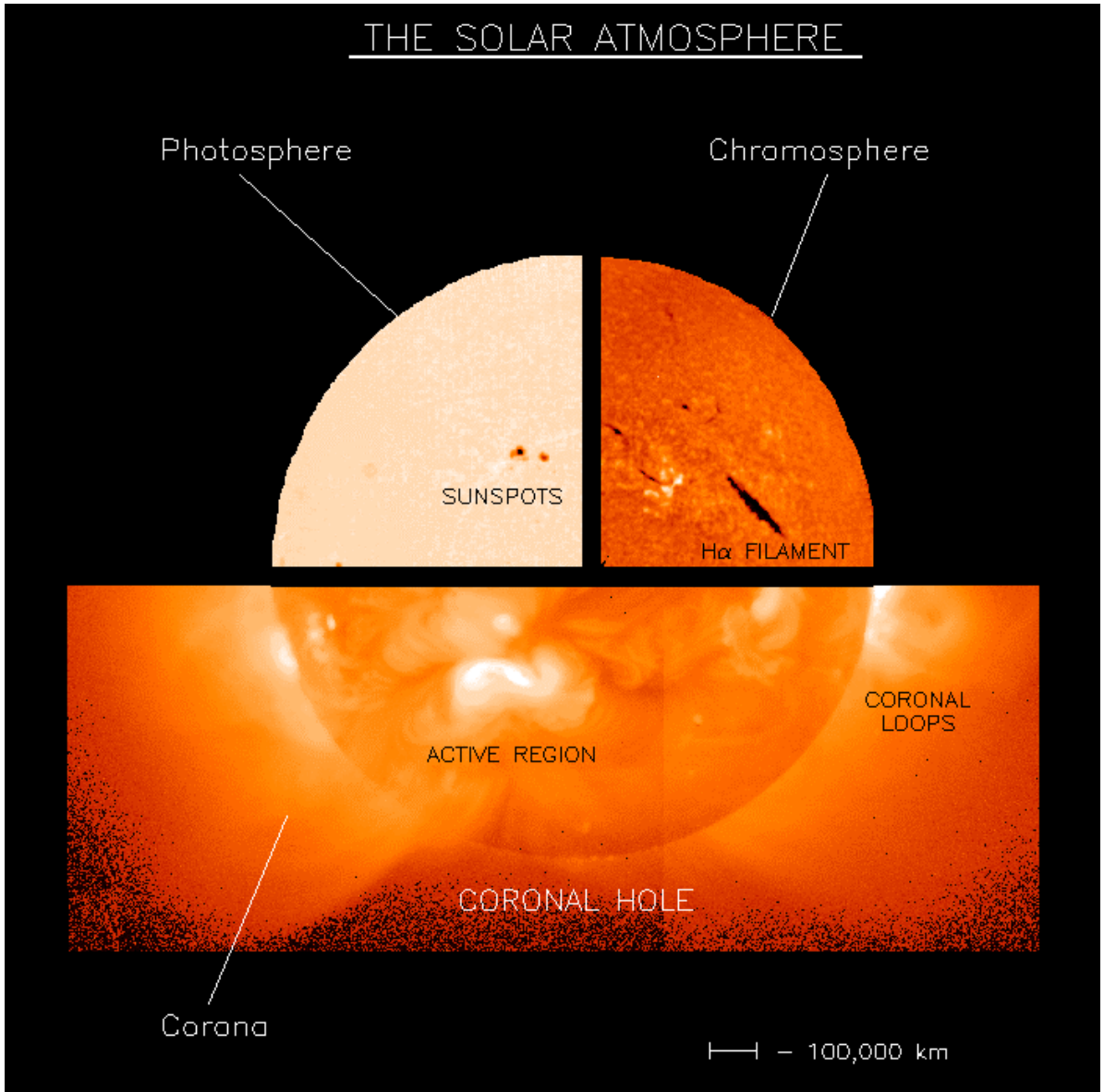
Nature of solar observations

- **For a complete picture of what is happening, we need to use as wide a range of observations as possible**
  - The appearance of the Sun changes dramatically with wavelength
  - Different layers of the solar atmosphere and material at different temperatures are best seen at different wavelengths

- **For technical and practical reasons:**
  - UV, EUV, X-rays and γ-rays observed from space
  - Radio and optical wavelengths observed from the ground
    - Issues related to coverage by each observatory
  - Differences in approach to handling data have developed

- **The observations used to build up a picture of the plasma in multi-dimensional parameter space (incl. *x*, *y*, *z*, *t*, *T* & ρ)**
  - How plasma contained in 3d structures evolves with time
  - Where and how energy released and how it affects the system
  - Etc…

Nature of solar observations



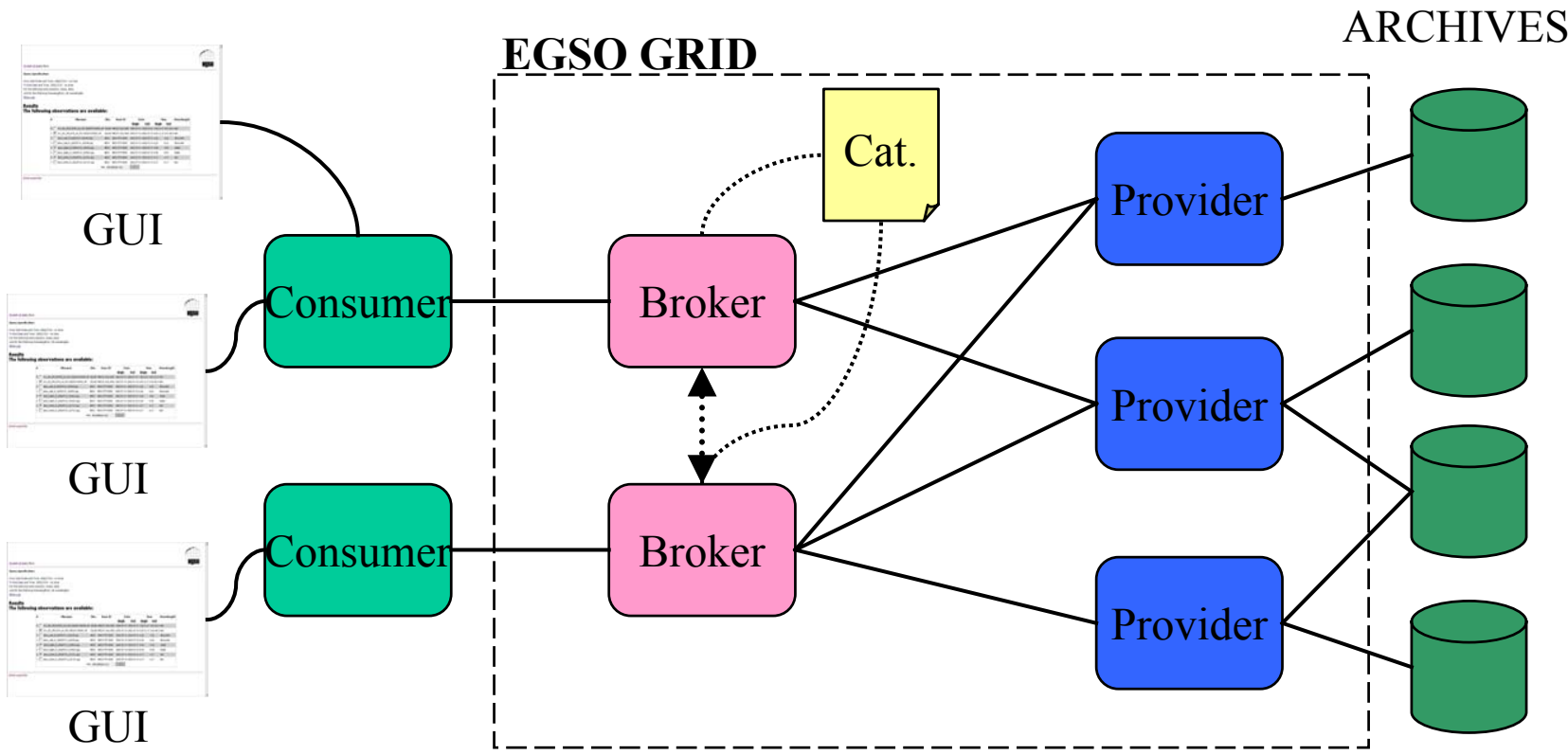
- Appearance of the Sun changes dramatically with wavelength
- For a complete picture of what is happening, we need to use as wide a range of observations as possible

Some generic issues

We need to build on the existing situation

- **User community scattered around the world**
  - Capabilities of users & their computing facilities vary greatly
  - Users want to know if data addressing a problem exists
    - Not really interested in where the data are located
    - Or, how the data are accessed, processed, etc.
  - Increasing desire for combined studies with other regimes
    - Astrophysics, Climate Physics, Space Weather, etc.
- **Data centres and observatories located around the world**
  - Large and small data providers (with varying resources)
  - Need to make it as easy as possible to add new data sets
    - Planned data volumes much larger than for current instruments
  - Cataloguing differs in quality, contents, and dependencies
  - Must handle multiple copies of data and proprietary data
  - Must ensure integrity of data providers
    - Authentication an issue that needs serious consideration
    - Need to minimize how it affects the user, etc.

Simplified Architecture



Architecture defined in terms of roles

Consumer, Broker and Provider

After R. Linsolas, IAS

Access to Resources

EGSO is a Grid and activities depend on access to resources

- **Resources described by entries in a Resource Registry and managed by a Broker. Types include:**
  - Metadata – from prime data providers
  - Data – from data centres, observatories, etc
  - Processing – simple, multi-instance processors, HPC(?)
  - Storage – cache space, on-line mass storage, etc.
  - Services – support of complex (meta)data products

  Note: Some providers can support multiple capabilities

- **The Broker allocates resources and controls:**
  - How much being requested of a particular provider
  - Processing of data & staging of results
    - Processing may be at different site to data provider

- **Broker & Registries replicated to provide system resilience and permit load sharing**

The EGSO Search Engine

In order to provide an enhanced search capability, EGSO will improve the quality and availability of metadata

- Enhanced cataloguing describes the data more fully
  - Standardized metadata versions of observing catalogues tie together the heterogeneous data sets
  - New types of catalogue allow searches on events, features and phenomena rather than just date & time, pointing, etc…
- Ancillary data used to provide additional search criteria
  - Images, time series, derived products, etc.
- Search Registry describes all metadata available for search
- It will be possible to access to EGSO through:
  - A flexible Graphic User Interface (GUI) – normal route
  - An Application Program Interface (API) – this provides access for users from other applications, communities or Grids

The enhanced solar catalogues

- **Unified Observing Catalogues (UOC)**
  - Metadata form of observing catalogues used to tie together the heterogeneous data, leaving the data unchanged
  - Self describing (e.g. XML), quantised by time and instrument, with no dependencies on *ancillary data* or proprietary software and any errors corrected
  - *Standards* defined for future data sets (e.g. STEREO, ILWS, Solar-B)
- **Solar Event Catalogues (SEC)**
  - Built from information contained in published lists
  - Flare lists, CME lists, lists in SGD, etc.
- **Solar Feature Catalogue (SFC)**
  - Lists of the occurrence of events, phenomena and features provides an alternate means of selecting data
  - Derived using image recognition software developed in WP5

Similar hierarchical cataloguing required in other data Grid projects

The enhanced solar catalogues

- **Unified Observing Catalogues (UOC)**
  - Metadata form of observing catalogues used to tie together the heterogeneous data, leaving the data unch
  - Self describing (e.g. XML), quantised h ment, with no dependencies on ancil oftware and any errors correct
  - *Standards d* lar-B)
- **Sol**
  - B
  - Fl
- **Solar F**
  - Lists events, phenomena and features prov mate means of selecting data
  - Derived using image recognition software developed in WP5

Objective of the improved metadata, etc. is to pose questions like:

Identify events when a filament eruption occurred within 30° of the north-west limb and there were good observations in Hα, EUV and soft X-rays

Similar hierarchical cataloguing required in other data Grid projects

Solar Event Catalogue (SEC) Server

- **Server specializing in event catalogues**
  - Complexity of updating information hidden from the system
  - RDBMS that can accept SQL queries
  - Interfaced into **EGSO** as a Web Service
  - Results returned in **VOTable** format
  - Test interface available through URL:
    http://radiosun.ts.astro.it/sec/sec_ui.php

- **Currently being tested with:**
  - Flare lists from NOAA and REHESSI
  - NOAA Proton Event list
  - CME lists from SOHO/LASCO
- **In process of being added:**
  - NOAA Active Region (NAR) database
  - Various indices – Sunspot Number (SSN), 10.7cm flux, Kp

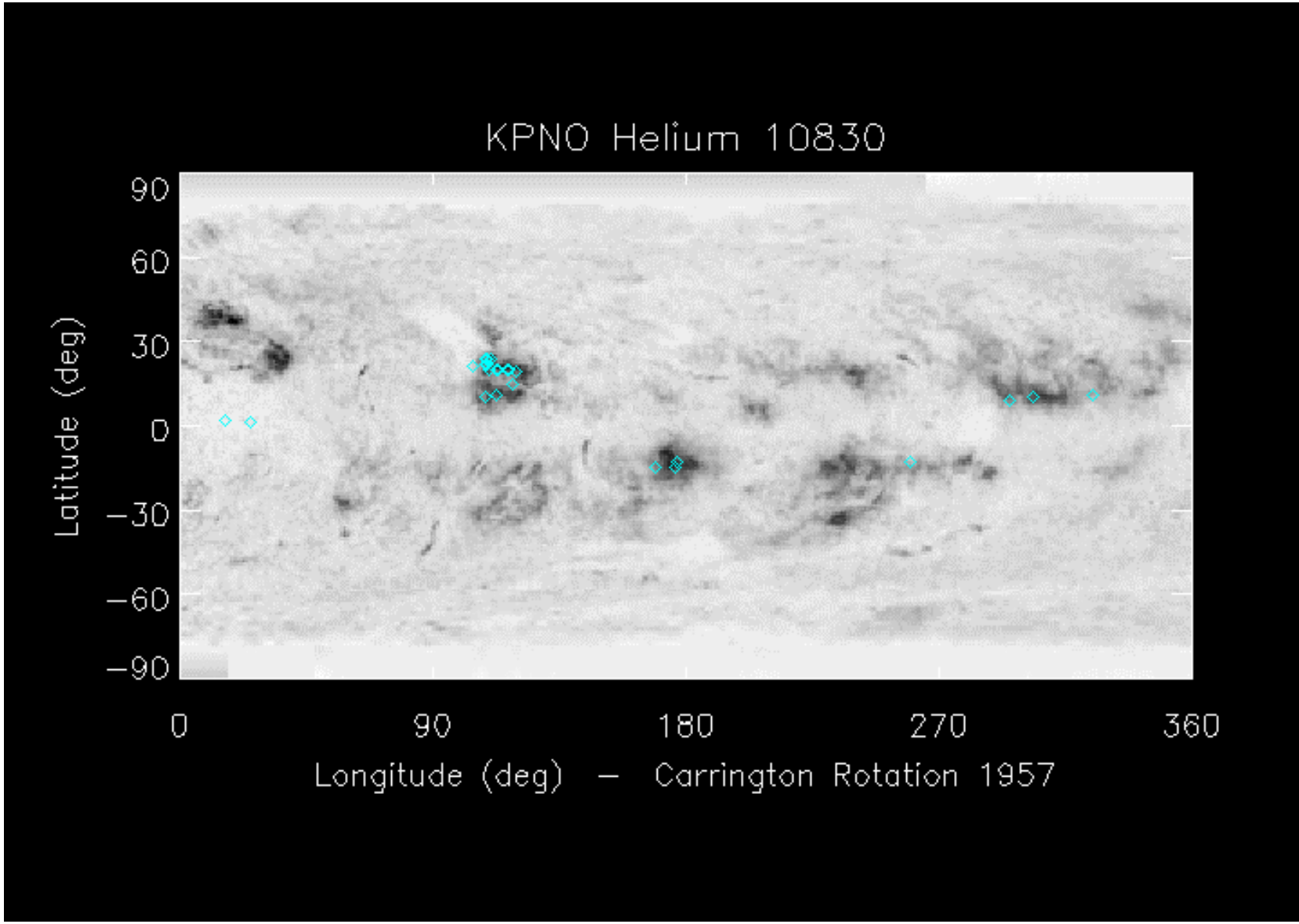
egso

Solar Event Catalogue

QL query: **select time_start,time_end,nar,latitude,longitude,xray_class from sgas_event WHERE time_start between \'1999-12-05 0:00:00\' AND \'2000-01-01 23:59:59\' order by time_start**

ime_start	time_end	nar	latitude	longitude	xray_class
999-12-06 07:00:00	1999-12-06 07:34:00	8788	10	43	C8.4
999-12-06 11:15:00	1999-12-06 11:17:00				
999-12-06 20:04:00	1999-12-06 20:05:00				
999-12-06 21:18:00	1999-12-06 21:26:00				C4.7
999-12-07 09:38:00	1999-12-07 09:39:00				
999-12-07 21:15:00	1999-12-07 21:23:00	8791	-13	66	M1.0
999-12-07 21:36:00	1999-12-07 21:36:00				
999-12-09 01:05:00	1999-12-09 01:05:00				
999-12-09 12:31:00	1999-12-09 12:34:00				
999-12-09 16:40:00	1999-12-09 16:46:00				
999-12-09 20:34:00	1999-12-09 20:34:00				
999-12-10 04:09:00	1999-12-10 04:17:00				C1.1

Search results over-plotted on synoptic map



Solar Feature Catalogue (SFC)

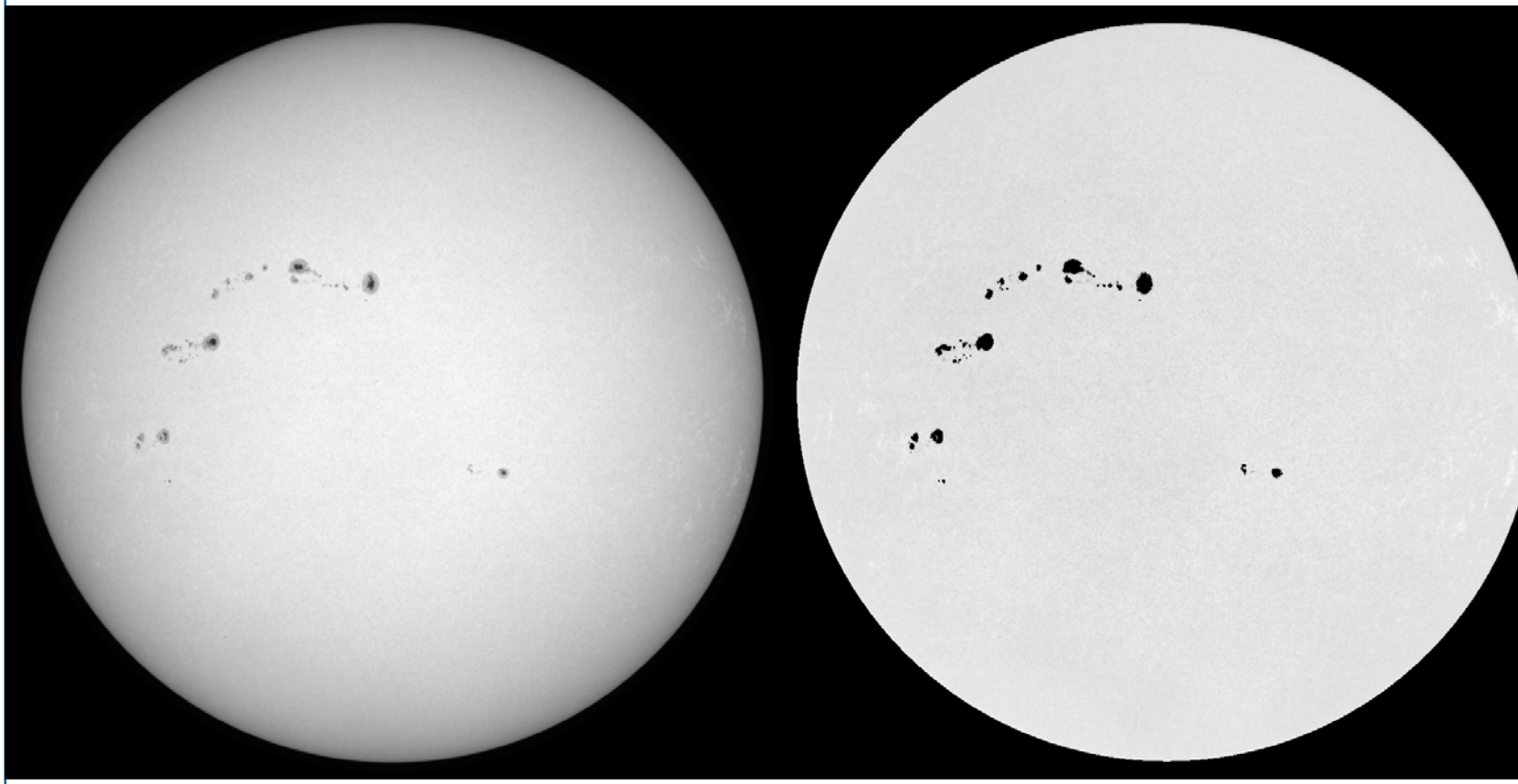
- EGSO has a work package (WP5) dedicated to developing tools needed to detect common solar features and then using them to derive the Solar Feature Catalogue

- Code to prepare images for recognition already available
  - Removal of artifacts, regularize shape, etc.
- Work in progress on codes to detect the features
  - Codes for sunspots, active regions and filaments developed and under test
  - Codes to recognize coronal holes and magnetic neutral lines under investigation

- Trying to define standard way of describing features for the feature catalogue
  - Preliminary version of SFC has been prepared
  - Now starting experimenting with the results to determine if objectives can be realized with the stored information

Sunspots detection in white light

Original image on the left and detected sunspots on the right



Use of the Solar Feature Catalog

- **The SFC can be used in at least three ways:**
  - Outline features recognized in one wavelength on an image taken in another (at a different time)
  - Determine when events related to features have occurred – e.g. filament eruptions, flux emergence
  - Track relative motion of features – e.g. sunspots

- **The SFC will be deployed as a Server addressed through Web Services**
  - Not clear whether the SFC Server will be combined with the already deployed SEC Server
  - Server will be accessible to other VO projects
  - Results will be returned in VOTable format in some cases; other formats TBD

Review of Metadata - I

- **Data sources are scattered over the globe**
  - Data are very heterogeneous with many different formats
  - Do not want to alter the data or make any particular requirements on the data providers
  - EGSO must be able to accept data in whatever form and means of access provided (http, ftp, Web Services…)
    - Requirement that all providers use GRID technology is not realistic
  - Catalogue information – i.e. the UOC – is key to providing a homogeneous means of accessing the data
    - UOC is in standardized format with dependencies removed…
- **Search Registry summarizes what observations have been made**
  - Abstraction of the UOC, at some granularity
  - Registry held centrally – exact location of UOC is to be clarified
  - Facilitates searches even if all providers not online
  - Allows search optimization because Registry records can be ordered as required – the data are not affected

Review of Metadata – II

- **EGSO will provide enhanced ways of searching for data**
  - Basic search uses Search Registry and UOC and is based on date & time, pointing and wavelength.
  - More sophisticate d searches use new metadata – the SEC and SFC – to search on events, features and phenomena
    - Eventually reduces to a search based on date and time, etc.

- **Finding the data is a different problem to determining if observations were made**
  - Data records in the Resource Registry describe what data are held at each provider
    - Many different types of access must be accommodated
  - Ensures EGSO directs data requests only to providers that are able to satisfy the request
  - Also allows balancing of loads, aggregation of resource poor providers, user preferences, etc.

Conclusions

- Of necessity the solar community is moving towards a virtual environment to access solar and related data
- EGSO is a Data/Computing GRID that will enhance access to solar data and provide advanced search capabilities
- EGSO is attempting to create a system that will encourage participation using GRID technology while not making strong requirements on the participants
  - We believe that this reflects reality in many application based GRID projects

- For more information on EGSO see:

  http://www.egso.org
- Or e-mail:

  bentley@egso.org