Active Brokerage for Data Grids

A Sanna and C Zunino† and B Bentley and G Piccinelli‡

† Politecnico di Torino, ‡ University Colloge London

Abstract: Scientific research and practical applications of solar physics require data and computational services to be integrated seamlessly and efficiently. The European Grid for Solar Observations (EGSO) leverages grid-oriented concepts and technology to provide a high-performance infrastructure for solar applications. In this paper, we describe the active brokerage technique adopted in EGSO to improve the access to and flow of solar data. The technique is based on a multi-tier management of metadata.

1 Introduction.

A major hurdle in many areas of solar research is locating data that effectively address the problem being studied and retrieving the identified datasets [2]. New ground- and space-based instrumentation will soon produce volumes of data that are not sustainable by many of the current approaches to data utilization. EGSO, the European Grid of Solar Observations, leverages new methods of resource federation, including Grid computing, to lay the foundations for a virtual solar observatory [3, 4, 8]. The main technical objectives for EGSO include: the federation of solar data archives, the creation of tools to select, process and retrieve distributed and heterogeneous solar data, the creation of tools to compile standardized catalogues for space and ground-based observations, and the creation of tools for the automatic generation of solar feature catalogues.

The ultimate goal of EGSO is to satisfy the complex information needs of solar physicists in a way that insulates them from issues such as location and distribution of data sources. Solar physicists will be provided with graphical interfaces similar to the prototype presented in Figure 1. Data of different types are interactively aggregated and manipulated based on the content logic of a solar experiment. The actual location of data sources and processing capabilities is managed automatically by the EGSO infrastructure. As an example, the picture in Figure 2 is a solar stack. The four solar images derive from four different types of data. From bottom to top: H-alpha emission, extreme ultraviolet, soft X-ray, and radio emission. The data sources are actually located in four different observatories. Solar physicists concentrate on parameters such as time interval and type of data. EGSO handles all the operations involved in the location of data sources, retrieval of the data, and data processing (e.g. calibration).

In this paper, we present an overview of the current version of EGSO. The focus is on the aspects and parts of the system more closely related to the management of flows of data and metadata. In particular, we outline the some of the solutions adopted in EGSO to leverage active brokerage.

2. User Requirements.

The information in this section represents an extract of a six-month user investigation done within EGSO. More information on the process followed as well as the outcome of the requirement gathering activity can be found in [7]. The extract is focused on data and metadata.

From user indications, the EGSO system must support a framework of metadata structures that incorporate all entities and attributes defined by the data models of the current solar physics archives. This framework needs also to incorporate entities and attributes defined by all non-data entities in the system. Specifically, it needs to include administrative, structural, and descriptive metadata. The framework must include semi-structured and incomplete data and metadata. The framework also needs to be extensible and support dynamic modifications and annotations. The EGSO system must incorporate new data as they are generated. The metadata framework should be able to incorporate new metadata as they emerge from data processing. The EGSO system should support translation between different metadata structures and correlate multiple data resources as required.

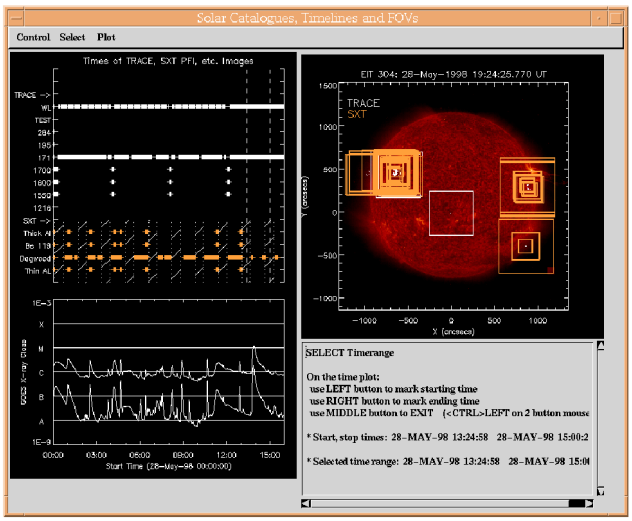

Figure 1: Interface for solar data manipulation

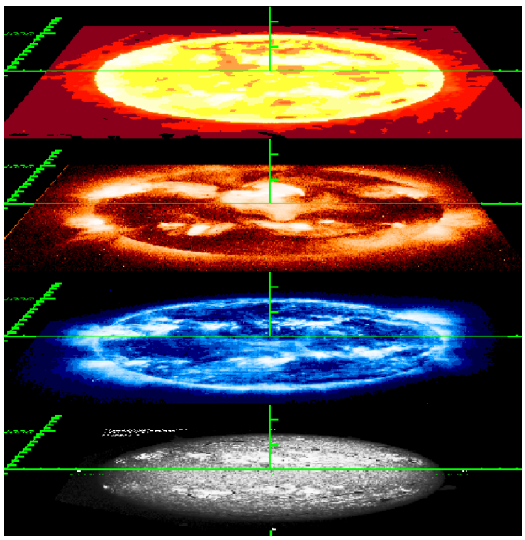

Figure 2: Views of the sun based on different types of data

Due to technical motivations as well as operational policies, the sharing of data and metadata can vary considerably between different organisations within the EGSO community. For example, some data providers are willing to distribute entire catalogues of the data that they have. Other data providers give instead just general indications, and they use catalogues internally to service specific requests. Along the same lines, some providers prefer to outsource most of the request screening. Other providers prefer instead to evaluate directly every data request they service. The EGSO system must be able to adapt to the sharing preferences of different users.

3. The EGSO Functional Architecture

The functional architecture for the EGSO system defines three distinct roles: consumers, providers, and brokers. An organisation can play multiple roles. A *consumer* interacts with a *broker* in order to get indications of the provider(s) likely to hold specific data (offer a specific service). The broker provides to the consumer references to providers as well as information that the consumer can use for the selection process. The broker can make the choice for the consumer if required. Multiple brokers can cooperate to satisfy a request, but the interaction between brokers is transparent to the consumers. The *consumer* interacts with the *provider(s)* in order to refine the data (service) requirements. In the end, the *provider* delivers the data (service) directly to the *consumer*.

A *provider* interacts with a *broker* in order to supply information about the data (services) available to consumers. The multiplicity of brokers is transparent to the providers. Provider and broker need to agree on the actual information about data and/or services that the provider has to supply to the broker. Provider and broker also agree on aspects such as ontology, data format, and upload process (e.g. frequency, initiator). Specific attention is paid to the evolution of properties of data/service descriptions. The provider can instruct the broker on the actual consumers that the data (services) are available to. The *provider* interacts with the *consumer(s)* in order to refine the data (service) requirements, and in the end it delivers the data (service) to the *consumer*. The provider (P) can verify if a consumer (C) has been referred to P by a valid broker (B). P can also ask B for information about C to be used in the P-to-C interaction process.

A *broker* (Figure 3) interacts with *providers* and *consumers* and maintains a profile of the parties. In particular, the *broker* gathers from *providers* information on the data (services) potentially available to consumers. From the *consumers*, the broker gathers mainly information that can be used to customise searches for data and services. In addition to information on sources of data and services, brokers need to manage information on the accessibility of different data items and services to different perspective consumers. Brokers have the capability to track interaction with consumers and providers. Still, the level of tracking is adaptable to needs and wishes of the user community. A Distinctive element of the brokers is that they interact among each other to create a single virtual broker. Support services (e.g. caching, logging, auditing, format transformation, workflow management) are modelled as standard

services. For example, a broker can also cache queries or query results. Such activity is modelled as a caching service, and the broker becomes also a provider.

A vertical segmentation [5] of the infrastructure that EGSO provides to enforce individual roles includes three main layers: connectors and adaptors, role-specific infrastructure, and user interface. The lower-level components in each layer are grouped into role-dependent and role-independent. Such distinction eliminates the need of redundant components in case the infrastructure for multiple roles is installed on a single system. Connectors and adaptors are used to gain access to the existing resources of a given organisation. Connectors are used to reach a given resource (e.g. a database). Adaptors are used to homogenise the format of the information exchange between a resource and the upper layers of the EGSO system. Connectors and adaptors need to address the peculiarities of a wide range of software systems. EGSO will provide a development framework for the creation of both connectors and adaptors. The role-specific part of the infrastructure includes tools and technologies that address the core needs of individual roles. For example, in the case of a broker the infrastructure provided by EGSO includes a federation-oriented engine for information retrieval. A flexible user interface is layered on top of all the components of the role-specific infrastructure. The interface comes as a framework with coarse-grain components that individual users can tune to specific needs.

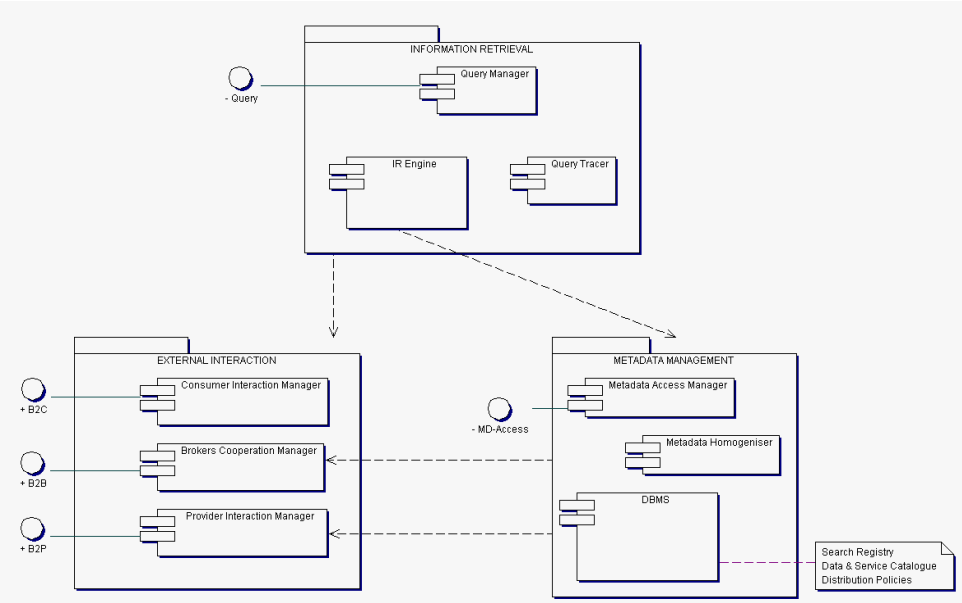

Figure 3: UML component diagram of an EGSO broker

4. The EGSO Broker

The two main subsystems within an EGSO broker (Figure 3) are dedicated to information retrieval and metadata management. Metadata mainly consist of catalogues of solar data that are produced and maintained by various observatories. Different observatories have different technical capabilities and operational policies. Also, different catalogues may be encoded in different formats. In response to user requirements, the EGSO brokerage system takes a flexible approach to metadata handling. Flexibility extends to both the physical partitioning of metadata and the interaction processes involved in accessing the metadata.

The metadata framework adopted by EGSO is based on a multi-tier approach. Given a catalogue, views can be created at different levels of precision and distributed between different parties. As an example, a provider P can retain a complete catalogue for a dataset. P can give a synoptic version of the catalogue to a broker B. In turn, B can give to all others EGSO brokers a condensed version of the synoptic catalogue. The retrieval algorithm used by EGSO brokers is based on a multi-hop technique similar to the one proposed by Ratnasamy [6]. The main extension to the basic technique is related to on-the-fly expansion of the search space. Within a search chain, the entity involved in the step N can ask for more information to the entities potentially involved in the step N+1. Hence the search space

can fluctuate. The entity at level N may subsequently need to dispose of the extra information acquired (e.g. because it is time-dependent). Moreover, one of the entities at level N+1 can refuse the extra information to the entity of level N. In such situation, the entity at level N can return to the entity at level N-1 either a "not found" result or a pointer to the entity at level N+1. The second option is applicable when the entity at level N+1 is prepared to give further details to higher-level entities. A typical example is the case in which a data provider requires direct interaction with the consumer for the refinement of a search (e.g. for tracking purposes). The EGSO brokerage system supports delegation at identity as well as action level. Hence brokers can operate as active negotiators on behalf of consumers in situations where information is not immediately available.

In addition to the retrieval of raw data, the active brokerage capabilities of the EGSO system encompass data and service composition. Composition refers to the possibility to satisfy individual requests by aggregating the results of multiple sub-requests. Composition can involve pure data, pure services, or a combination of data and service. As an example, composition of pure data can be applied to cover time intervals. The data related to different segments in the time interval can be available from different observatories (e.g. due to the rotation of the Earth). An example of data and service composition is outsourced calibration. The consumer can require the calibrated version of data that are stored only in uncalibrated format. The broker can locate a calibration service and that can process the raw data.

5. Middleware Platform

A number of technologies [1] were considered for the implementation of the EGSO brokerage system. Peer-to-peer solutions such as SUN's JXTA were compared against more service-oriented solutions such us Globus. The convergence of Globus and Web services into the OGSA (Open Grid Service Architecture) was the main motivation for the selection of Web services as reference platform. The part of the system more dependent on middleware (Figure 3) is the subsystem for external interaction. As standards and technology are in constant evolution, an evaluation of aspects of the system such as performance is premature. Nevertheless, early experiences show a positive trend towards product-level quality.

6. Conclusions

Grid-based solutions enable data and services to be exchanged efficiently across networks. Still, consumers and providers need to be suitably matched in order to maximise results and reduce waste of resources. The trial-and-error approach supported by Internet search engines is not applicable in many application domains. In the case of solar physics, requests not properly handled can cause hour-long processing of data catalogues and gigabytes of data unnecessarily sent across networks. Moreover, access and usage policies normally apply to sensitive data and high-value processing resources. The active brokerage techniques proposed by EGSO leverage multi-tier definitions of metadata to enable negotiation-based access to resources, as well as composition of data and services.

References

[1] L. Ciminera, A. Sanna, C. Zunino, N. Linsols, and G. Piccinelli *"Survey on grid and peer-to-peer network technologies"* EGSO Report EGSO-DE01_01-D03-021022, 2002.

[2] A. Csillaghy, D.M. Zarro, and S.L. Freeland *"Steps towards a virtual solar observatory"* IEEE Signal Processing Magazine, N.18/2, 2001 – pp.41-48.

[3] I. Foster, C. Kesselman, J. Nick, and S. Tuecke *"Grid Services for Distributed System Integration"* Computer, 35(6), 2002.

[4] F. Fox and D. Gannon *"Computational grids"* Computing in Science & Engineering, Vol.3 No. 4, 2001.

[5] ISO/IEC. RM-ODP. "*Reference Model for Open Distributed Processing*" Rec. ISO/IEC 10746-1 to 10746-4, ITU-T X.901 to X.904, ISO, 1997.

[6] S. Ratnasamy, P. Francis, M. Handley, R. Karp, and S. Shenker *"A Scalable Content-Addressable Network"* Proc. ACM SIGCOMM, 2001.

[7] K. Reardon, S. Giordano, and E. Antonucci *"User and science requirements document"* EGSO Report EGSO-WP1-D2-20021031, 2002.

[8] H. Stockinger *"Distributed Database Management Systems and the Data Grid"* Proc. 18th IEEE Symposium on Mass Storage Systems, 2001.