COLLABORA TIVE SOLAR SYSTEM RESEARCH DATA MANIPULATION

	Name	*Date*	*Signature*
Prepared by :	**C. Perry** (RAL)	10/05/2002	
Reviewed by :	**C. Perry** (RAL)	10/05/2002	
Approved by :	**S. Beco** (DATAMAT)	10/05/2002	
Accepted by:	**Gerhard Kreiner** (ESA) **Pier Giorgio Marchetti** (ESA)		

Distribution List

Company	*Name*	*Number of copies*
ESA	Pier Giorgio Marchetti	e-mail
	Gerhard Kreiner	e-mail
DATAMAT	Stefano Beco	e-mail
	Fiorella Lamberti	e-mail
ASPI	Veronique Perrin	e-mail
CSSI	Pierre Vielcanet	e-mail
RAL	Christopher Perry	e-mail
QIN	Fan Lei	e-mail
	David Rodgers	e-mail
SSL	Conrad Morris	e-mail

Document Status Sheet

Issue	Date	Details
1.0-D	05/04/2001	First Draft
1.1-D	10/05/2002	Updated draft after URR

Document Change Record

Issue	RID No.	Details
1.1	UR-SYS-SB-001	Page 1: "Accepted by:" field added and other fields re-organised to take account of this.
1.1	UR-SYS-SB-002	Page 2: Distribution list updated.

Table of Contents

List of Figures

1. INTRODUCTION

The aim of SpaceGRID in the Solar System Research (SSR) domain is to remove the barriers that currently limit the efficient analysis of distributed data by providing a framework for uniform access to, and manipulation of, local and distributed data sources. The SpaceGRID should not only endeavour to enable research that could not be done before. It should also facilitate analyses that up to now have been very difficult, but not impossible, by reason of data or information inaccessibility, or have only been possible by a chosen few by reason of resource limitations. The SSR part of SpaceGRID is, therefore, not only about handling large amounts of data or providing large scale high performance computing, but is equally about facilitating the handling of small but complex and diverse amounts of data. In the short-term the federation of existing SSR resources will result in improved and faster science return. Longer-term SpaceGRID will encourage the standardisation of design, promoting re-use of analysis systems, long-term security of data and reduced overall cost.

1.1 Purpose of the Document

Task 5 of [SPCGRD-SOW] called on the SpaceGRID consortium to establish the requirements, assess the benefits and evaluate the implementation options for ESA development of a collaborative data analysis system for the Solar System Research community. The purpose of this document is to provide an overview of collaboration facilities within the SSR domain.

1.2 Definition, Acronyms and Abbreviations

1.2.1 Acronyms

API	Application Programming Interface
CDF	Common Data Format
CD-ROM	Compact Disc – Read Only Memory
DBMS	Data Base Management System
ESA	European Space Agency
ESTEC	ESA Technology Centre
EU	European Union
FITS	Flexible Image Transport System
FTP	File Transfer Protocol
HPC	High Performance Computing
IDL	Interactive Data Language
ISTP	International Solar Terrestrial Physics
OGSA	Open Grid Service Architecture
OODT	Object Oriented Data Technology
PDS	Planetary Data System
RAID	Redundant Array of Independent Disks
RAL	Rutherford Appleton Laboratory
SCT	Special Conditions of Tender
SCSI	Small Computer Systems Interface
SPCGRD	SpaceGRID
SSL	Science Systems Limited
SSR	Solar System Research
STP	Solar Terrestrial Physics

STPDF	STP Data Facility
UK	United Kingdom
US/USA	United States
WP	Work Package
WWW	World Wide Web
XDF	eXtensible Data Format
XML	eXtensible Markup Language

1.3 REFERENCES

1.3.1 Applicable Documents

[SPCGRD-SCT] AO/1-3863/01/I-LG, Invitation To Tender – Study of GRIDS and Collaborative Environments for Space Application – Appendix 3 – Special Conditions of Tender

[SPCGRD -SOW] AO/1-3863/01/I-LG, Invitation To Tender – Study of GRIDS and Collaborative Environments for Space Application – Appendix 1 - Statement of Work

[SPCGRD-PROP] P-A876/USA/PT-656-01, Is. 1.1, Proposal for "Study of GRIDS and Collaborative Environments for Space Application (SpaceGRID)"

[PSS5-SP] BSSC(96)2 Issue 1, Guide to applying the ESA software engineering standards to small software projects

1.3.2 Reference Documents

[GRID] I. Foster and C. Kesselman, Eds., The GRID Blueprint for a New Computing Infrastructure, Morgan Kaufmann, 1999

[SPCGRD-URSR] Solar System Research User Requirements Document, SGD-SSR-RAL-REQ-001, April 2002.

[SPCGRD-SSSR] Solar System Research System Specification Technical Note, SGD-SSR-RAL-SSTN-005, April 2002.

[CSDSNTFR] Cluster II Project, CSDS Network Task Force Report, CL-EST-RP-2025, February 1999.

[APACHE] http://www.apache.org/

[ASTROGRID] http://www.astrogrid.org/

[CDAWeb] http://cdaweb.gsfc.nasa.gov/

[CSDS] http://sci2.estec.esa.nl/cluster/csds/csds.html

[EGSO] http://www.mssl.ucl.ac.uk/grid/egso/

[GBDC] http://www.wdc.rl.ac.uk/gbdc/gbdc.html

[GEEKLOG] http://geeklog.sourceforge.net/

[IDFS] http://www.idfs.org/

[ISDAT] http://space.irfu.se/isdat/

[OGSA] http://www.globus.org/ogsa

[OODT] http://oodt.jpl.nasa.gov/

[OPENBB] http://www.openbb.net/

[OVT] http://ovt.irfu.se/

[PAPCO] http://leadbelly.lanl.gov/ccr/software/papco/papco.html

[PDS-MOC] http://ida.wr.usgs.gov/

[SEC] http://www.sec.noaa.gov/

[SOHO] http://sohowww.nascom.nasa.gov/

[SOLARSOFT] http://sohowww.nascom.nasa.gov/solarsoft/

[SPARC] http://intel.si.umich.edu/sparc/

[SPICE] http://pds.jpl.nasa.gov/naif.html

[SPOF] http://www-spof.gsfc.nasa.gov/

[SSC] http://sscweb.gsfc.nasa.gov/

[TWIKI] http://twiki.org/

1.4 OVERVIEW OF THE DOCUMENT

The structure of this document is as follows.

- Section 2 provides a general description of collaboration within the SSR domain.
- Section 3 describes the facilities and tools that are currently used to enable collaborations, looking both at general and domain specific facilities.

2. GENERAL DESCRIPTION

Solar System Research (SSR) is a broad, multi-disciplinary science involving the study of a complex 3-d environment with interactions and phenomena that occur over a variety of temporal and spatial scales. Even when split into specific areas within the sub-domains, no single group or institution holds all the information required to adequately analyse, model and ultimately explain the processes involved. For example, in Solar physics SOHO data may be used in conjunction with data from other missions such as Yohkoh, TRACE or ground-based observatories. In Planetary studies SMART-1 data will need to be compared to data from previous lunar missions such as Apollo, Clementine and Lunar Prospector. In STP the problem is particularly acute due to the large number of multi-point in-situ measurements that are required to constrain the context of an event. Collaboration is therefore a vital part of the Solar System Research activity and a major part of all collaborations is the interoperability of the different resources required for a particular analysis. The importance of collaborations is seen in Table 1 (from the SSR user requirements analysis [SPCGRD-URSR]) that shows the vast majority of the community are involved in combining different datasets as part of their research.

Combining Data	
From the same mission	86%
From missions in same area	86%
With other activities within SSR	23%
Interdisciplinary (other domains)	7%

Table 1: Scope of inter-working with different data sets.

The collaborations are between groups that are geographically widely spread. However, the numbers involved in each research investigation are often quite small, typically involving a core of two to four individuals with each being responsible for the provision and interpretation of a particular data set or model.

The sorts of collaborations that are undertaken depend on the type of observation or analysis involved and can be categorised as follows.

- **Plan->observe->analyse.** In this case observations are planned in advance with collaborations being formed early in the process. The planning defines the optimal modes for the instruments involved and coordinates the timing of the observation to maximise the likelihood of accumulating good data. This technique is most commonly used with instrumentation that requires particular configuration or has particular constraints such as pointing or observation time. For example, ground-based radars may need to be conjugate with in-situ space measurements; or the fields of view of instruments on a planetary orbiter need to include the object of interest. The pre-planned observation clearly works best for static features or events that can be predicted reasonably well in advance.
- **Observe->search/identify->analyse.** In this case the search and identify stage takes place after the data has been accumulated. The techniques used to search a dataset will depend on the types of feature being investigated. They may involve using an activity index to locate periods of interest, or visually scanning a set of survey plots for a particular signature. Once a set of events has been identified, the datasets required for the analysis

must be collected together. This is a time consuming job requiring many interactions with different groups and online resources. It is normally at this stage that collaborations are formed or events rejected because the required data are not readily available (e.g. because the required instrument was in the wrong mode). This method works well for data where there is near continuous coverage and where features are difficult to predict. However, it can lead to selection effects since effort is concentrated on the most prominent, and therefore easiest to spot, events.

- **Observe->statistical analysis.** This is similar to previous case but involves the systematic analysis of a large number of observations, or long periods of data, to determine some overall property. This avoids some of the selection effects that can result from individual event identification. However, the large volumes of data that must be handled mean that these techniques are currently only applied to simple data sets available locally to the researcher.

Reduction of the raw instrument data requires detailed knowledge of the operation and limitations of the instruments or models used to generate it. Within collaborations, individuals are responsible for the provision and interpretation of the high quality data from their own instruments or models. The detailed analysis that is applied to these data will vary from study to study depending on which aspects of the data are important. However, one of the most critical tasks is the accurate combining of the different datasets, to allow comparison of observations taken under different conditions, or with the models and simulations that play a central role in understanding the processes involved. This may involve joining time series data onto a common timeline or re-sampling images so that they can be overlaid. In all case the task can only be automated if the data is supplied with a well defined and accurate set of metadata.

The sorts of data that are used within a particular collaboration can be categorised as follows.

- **Core data.** These are high quality data that have been fully reduced and interpreted, based on specific instrumental knowledge and underpin the particular study.
- **Ancillary and support data**. These are used as part of the analysis of the core data set, usually in standard ways that that do not require specialised interpretation. For example, the use of magnetic field information to order particle measurement into pitch angle distributions; or the use of solar wind velocity measured at L1 to calculate the propagation delay from the Sun to the Earth.
- **Contextual data**. These are used to provide a broader view of the system being analysed by the core measurements. Data maybe lower quality than core measurements and is not used as part of the detailed analysis. For example, Viking images may be used to provide the context for higher resolution images from Mars Global Surveyor; or the direction of the interplanetary magnetic field maybe used to indicate the coupling between the solar wind and the magnetosphere. Supporting data from other instruments or spacecraft.
- **Auxiliary data**. These are required for all of the above, and include items such as orbit and planning information, events and constraints.

In each case comprehensive metadata is required in order to allow the data to be used in a systematic manner without the need for time consuming transformation and manipulation of the data at the individual parameter or record level. Without these metadata it is very difficult for collaborations to proceed in an efficient manner.

3. COLLABORATION FACILITIES AND TOOLS

Since the numbers involved in collaborations are relatively small, the favoured methods for communication are via E-mail, workshops and one-to-one phone conversations. The use of concurrent collaborations methods is not widespread as shown by Table 2, taken from the user requirements analysis [SPCGRD-URSR]. Even teleconferencing, which is fairly widely available, is only used by a minority of existing researchers.

Communication methods	
Tele-conference	46%
Video-conference	10%
Net meeting	4%
AccessGRID	0%

Table 2: Concurrent collaborative communication methods

Instead, collaborations are currently the result of offline analysis with subsequent distribution of processed data products using local web sites and ftp servers. To some extent this is due the lack of availability of good concurrent analysis tools and a need for researchers to change the way in which they work to benefit from these relatively new facilities. However, as was seen from the user survey, the real benefit to collaborative working is from improved interoperability between the currently heterogeneous and widely distributed resources.

The following sections present some of the existing tools that are available to support research collaborations.

3.1 GENERAL

The most widespread tool for e-collaboration is E-mail. This may seem obvious but is easily overlooked. It provides a simple, yet effective, method for communication both between individuals and groups of researchers. It has the advantage that it is readily accessible and so no additional effort is required for its use. The availability of distribution lists and listserv facilities make it a convenient way to contact an extended community, and improved handling of embedded content and attachments has made E-mail a viable means for the exchange of processed results, such as plots and small volumes of data. However, it becomes less efficient as the volume of products and the numbers of collaborators increases, resulting in a deluge of uncoordinated information, only some of which may be relevant to individual recipients.

There are a number of alternatives to E-mail including news facilities [GEEKLOG], bulletin boards [OPENBB] and collaborative document sharing facilities [TWIKI]. These are being very successfully used to support the technical and science specification of the [ASTROGRID] project. They have the significant advantage that information can be split into appropriate topics and threads, and tools are provided to assist in the navigation and location of particular items. They also have the benefit that the information is centralised so that it can be easily archived and only the parts that are of interest to the user need be downloaded. However, this centralisation is also an obstacle to the widespread use of these facilities, since it requires effort from one of the participating groups to installed and maintain the system.

3.2 Domain Specific

In the following sections we consider existing support within the sub-domains for collaborations and data interoperability.

3.2.1 Planetary

Within the planetary sub-domain most data is available via the US Planetary Data System. One of the main aims of the PDS is to ensure the long-term availability of high quality planetary data. This is done through comprehensive documentation of the products supplied to the system, the use of standardised PDS label describing the data and formal review of the datasets. While the diverse nature of the underlying data means that there is no single software package that can be used to manipulate all PDS data. The standardisation in descriptions means that development of standalone analysis routines for use within a particular collaborative research venture are greatly simplified. This capability is further enhanced by the use of a common mechanism for the specification of auxiliary data. In particular the use of [SPICE] kernels for the description of orbit, attitude, and instrument geometry information.

The PDS also provides a variety of online mission specific facilities to provide easy access for research groups wanting access to the stored data. For example the PDS Mars Orbiter Camera image collection [PDS-MOC] (Figure 1) provides online access to high-resolution images from Mars Global Surveyor together with lower resolution contextual images from the Viking mission. These can be used to survey the data set prior to downloading the full PDS data.

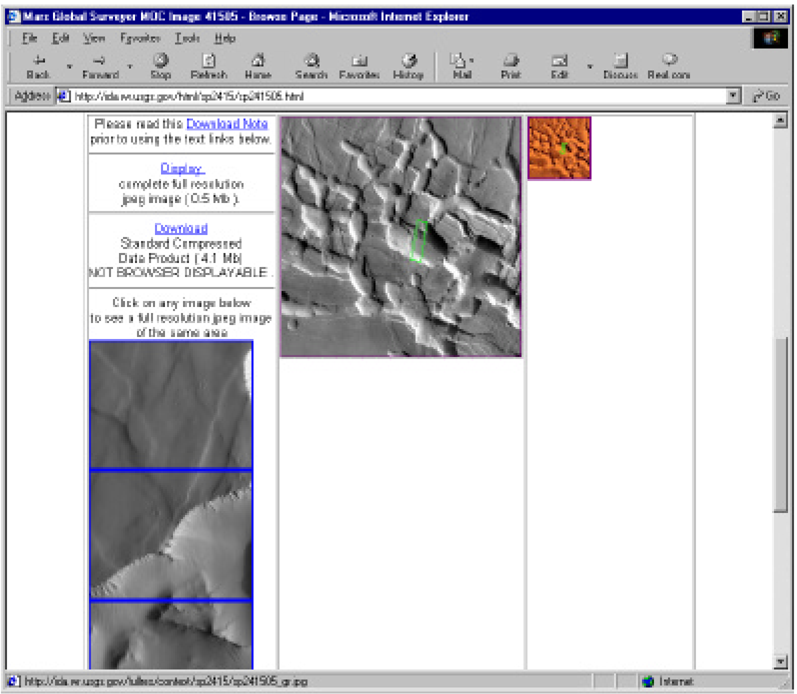


Figure 1 Mars Orbiter Camera collection together with contextual Viking images.

The PDS are actively involved in the development of facilities for improved data location and interoperability. These initiatives include the extension of the existing Distributed Inventory Tracking and Data Ordering System (DITDOS) that provides access to distributed data services and datasets. It uses a hierarchical set of profiles describing the resources at each level. Enhanced data interoperability is being developed as part of the Object Oriented Data Technology project. This uses a combination of XML and CORBA to pass standardised data and metatdata over the network. The first public implementation of the [OODT] will be for a selection of instruments from the Mars Odyssey mission and is expected to become available in the autumn of 2002.

3.2.2 Solar

Collaboration within the solar sub-domain is promoted through the use of a set of integrated software libraries, databases and system utilities, primarily running under IDL, called SolarSoft (see [SOLARSOFT]). The system provides solar physicists with a common data analysis and research environment that can handle data from a large number of existing sources (e.g. SOHO, TRACE and Yohkoh). Standard utilities are provided to handle common tasks such as time series analysis, coordinate transformations, spectral fitting and file manipulation. Mission and instrument dependent functions are supported through a hierarchical source code distribution that allows only the required support to be installed and simplifies the update of existing functions or addition of new instrument processing algorithms. The primary goals stated by [SOLARSOFT] are:-

- Provision of a large reusable software library
- Provision of a system that is largely hardware and site independent
- Promotion of standards to facilitate the coordinate analysis of solar data
- Promotion of an evolutionary environment
- Provision of access to supporting ancillary databases
- Provision of a file format independent analysis environment
- Provision of integrated access to other IDL packages.

Given widespread use of the system by the solar community, it appears that the developers have been successful in meeting their goals.

The main bottleneck in the collaborative solar analysis is therefore not in the processing of the data but in the identification, location and retrieval of interesting features or events. Survey images from sets of instruments are accessible from various online sources (e.g. SOHO, Figure 2 and ground based images from SEC, Figure 3).

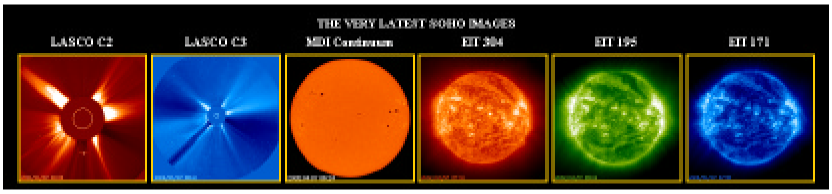


Figure 2 [SOHO] preview images from three instruments LASCO, MDI and EIT.

However, survey plots are only provided for a subset of all available data. When it actually comes to locating or requesting data many of the existing systems are limited to selection of files by parameter (date, position, wavelength etc) with little or no information for the user on the characteristics of data to assist with making a sensible selection. The large volumes of some of these data have required the use of tape storage robots. Requests for data may be submitted online but the user must then await an e-mail notification of the data being available online or delivery of hard media. Once notified the user must then retrieve the selected data via ftp and only at this stage find if it is really useful. Data files are generally FITS or binary but contents differ (e.g.

some have single observation per file, some multiple images per file or even multiple files per image).

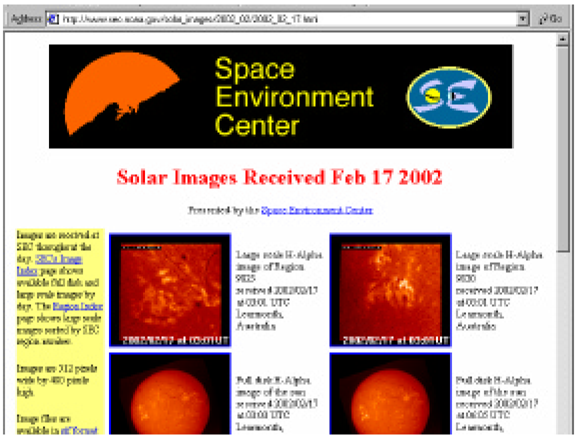


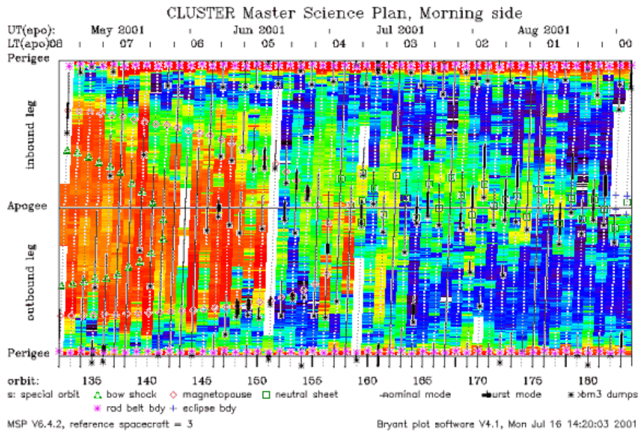
Figure 3 Preview images from ground based solar observatories provided by [SEC]

The provision of improved observing and solar feature catalogues are two of the main objectives of the EU European GRID of Solar Observations project [EGSO]. These will be a significant aid in the process of locating all the data sets required for collaborative work. The EGSO project is in partnership with both SpaceGRID and the UK AstroGrid project, with the aim of providing good interoperability between these different GRID initiatives.

3.2.3 STP

The need for collaborative working is particularly important in the case of STP where a large number of multi-point in-situ measurements are required, both for the core analysis, and to provide the context of an event. A group of researchers wishing to study the propagation and effect of a Coronal Mass Ejection event on the near-Earth environment may use data from SOHO, ACE, Cluster, Polar, IMAGE, ground based radars, geomagnetic activity indices to name but a few, and comparing them with outputs from models and simulations. Such analyses are extremely laborious involving the location of data from various archives and online resources. Due to changing calibrations or improved algorithms it may well be an iterative process. In addition suitable metadata must be obtained, which will almost certainly require the involvement of a member of the instrument team in the collaboration.

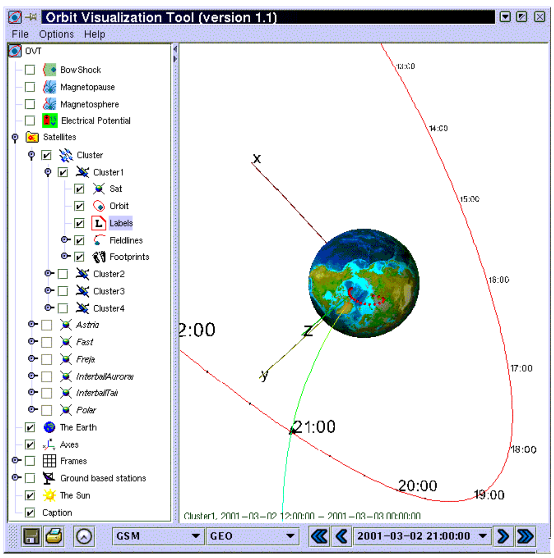
Because of the difficulty in predicting the dynamics of the near-Earth system, event analysis relies heavily on serendipitous measurements from spacecraft that happened to be in the right place at the right time. The identification of such events is either based on geomagnetic activity indices such as Dst or Kp, or on visual inspection of survey plots provided by the mission. To allow these visual searches to be performed rapidly the survey plots should be pre-generated so that they are instantly available and preferably include data from multiple instruments. NASA's Spacecraft Payload Operation Facility [SPOF] provides online survey plots for several of the current ISTP missions and a similar facility is provided for the Cluster mission by the [CSDS] (Figure 4) as well as by many individual instrument teams. However these individual facilities are not well coordinated so that multiple sites may need to be accessed in order to assess the quality of a particular event.



Error! Unknown switch argument.

Figure 4 Quicklook plots from the ESA Cluster Science Data System

In order to identify periods likely to be suitable for collaboration using data from multiple spacecraft, good knowledge of their locations is required. A key online facility in this respect is the NASA Satellite Situation Centre [SSC] that provides predicted and reconstituted orbit information for more than 60 spacecraft. The SSC also provides a set of online plotting (Figure 6), mapping and query tools to assist with the identification of good conjunctions between different spacecraft or with ground-based facilities. The Orbit Visualisation Tool [OVT] is available as either an online tool or an application that can be downloaded and run locally. It provides advanced capabilities for interactive visualisation (Figure 6) of spacecraft orbits allowing the user to focus in on particular areas of interest for example to assess the expected order in which the four Cluster spacecraft will cross the magnetopause or bow shock.



Error! Unknown switch argument.

Figure 6 Spacecraft orbit displays as provided by SSC & CDAWeb (left) and OVT (right)

An online facility directly supporting online concurrent collaborations using multiple datasets is Space Physics and Aeronomy Research Collaboratory [SPARC] (formally UARC). This system maybe operated either as an online workshop environment including concurrent chat facilities (Figure 8) or users can configure private views of the available data sources for individual investigations.

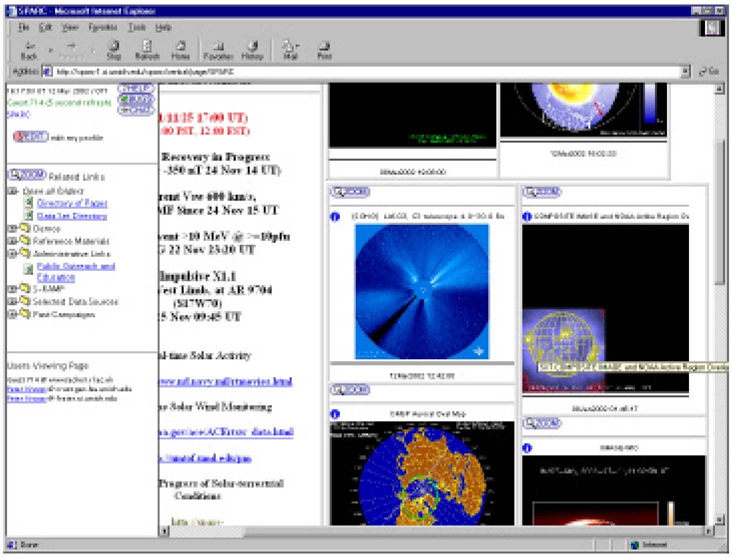
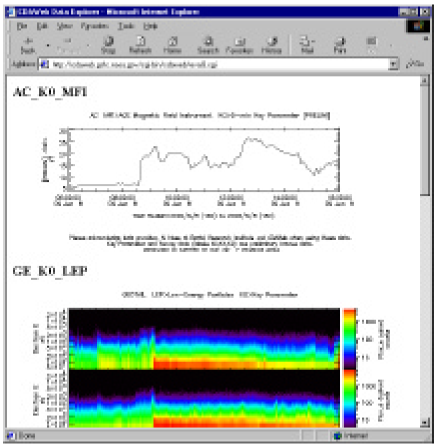


Figure 8 The Space Physics and Aeronomy Research Collaboratory

The SPARC system is a portal providing a common point of access to online plots generated by a number of participating facilities including data archives and modelling groups. The use of plots rather than a common underlying data infrastructure has the advantage that sophisticated plots provided by the individual facilities can be displayed. However, it has the disadvantage that the data is not directly available to the user to manipulate, combine or use locally. A similar but less sophisticated system, excluding the concurrent collaboration facilities, is the Cluster Ground Based Data Centre [GBDC] summary plot system. This provides a portal to online survey plots from more than thirty facilities, stacking and scaling the heterogeneous plots to aid inter-comparison of these time series data.

Once suitable events have been identified using the facilities described above the main issues for collaborative working are over data location, access and standardisation. Data location is made difficult by the widely distributed nature of the large number of in-situ and ground based measurements that are available. While the existing portal sites provide common access to survey plots, the data is often not as readily accessible. A common starting point will be the NASA Coordinate Data Analysis Web [CDAWeb] that provides survey level data from most ISTP missions, all stored in the Common Data Format (CDF) and including a good level of metadata description. The system provides facilities for selecting, previewing, generating ASCII listings and downloading data files (Figure 9). CDAWeb is probably the single facility that has had the largest impact on promoting collaborations within STP. One of the main limitations is that majority of available data is of the survey “key parameter” level and therefore only suitable for preliminary research activities. For anything other than these preliminary studies, high quality products are needed and these are not generally available from a central source, instead requiring direct interaction with the individual instrument teams resulting in provision of data in a variety of heterogeneous formats.



Error! Unknown switch argument.

Figure 9 Example quicklook plots from the NASA Coordinate Data Analysis Web

In addition to the online web-based facilities described above there are also a number of installed systems that although not widely used by the whole community do provide support for collaborative working within significant groups.

- Instrument Data File Set [IDFS] and associated science analysis system SDDAS provides a system for the description of low-level data files. Calibration factors are applied when the data is used allowing for changes in calibration factors to be applied without regenerating the underlying archive. The system supports a data promotion system allowing collaborators to access just the products that they require from the archive sites.
- [ISDAT] is a database management system for scientific data. It uses a client server model to provide access to a data held at different sites. The data servers are kept fairly simple to ease the extension of the system to support new facilities and data sets. Standard clients are provided to control data access, plotting and data analysis. More complex functions can be added through the development of specialised client applications
- [PAPCO] the Panel Plot Composer is an IDL application designed to provide multi-instrument plots from various ISTP missions including Polar, Cluster/RAPID, LANL, GPS and CRRES. Each data set has a set of functions that are used to load and manipulate the data somewhat reminiscent of SolarSoft described earlier. PAPCO has some limited ability to collect data provided by remote groups by using the GNU command line web system, wget, to retrieve files conforming to a pre defined naming convention.

Some work has been done on interoperability between the IDFS and ISDAT systems, details of which can be found at http://terezka.ufa.cas.cz/data/idfs2isdat/index.html