DATA PROCESSING, DATA BASES AND TOOLS AVAILABLE: PRESENT AND FUTURE

D. Heynderickx, B. Quaghebeur

Belgian Institute for Space Aeronomy, Brussels, Belgium

ABSTRACT

By their nature, space weather applications rely heavily on access to and processing of data collected by satellite or ground based instruments. Consequently, the availability of the relevant data over computer networks and powerful processing tools are vital for building and maintaining space weather applications, such as: describing the current state of the magnetosphere; prediction of "events" that change the environment and may interfere with operations of spacecraft and ground based systems; reliable warnings for operators and astronauts.

Data relevant for space weather applications include: solar activity, solar wind parameters, magnetospheric particle fluxes, magnetic field measurements, magnetic indices, auroral images. These data are continually collected by a large array of spacecraft and ground based observatories, and made available to the community as on-line data or via mass storage distribution.

Space weather applications use data from different sources (typically, different satellite missions), which have to be combined in order to study the effects on systems. Combining data sets is greatly simplified when a common storage format is used, so that analysis tools can be applied to data irrespective of their source and content, through a common interface. Two systems of data formats, besides storage in native format, are now in common use: NSSDC's Common Data Format (CDF) and SwRI's Instrument Data File Set (IDFS). On these formats, several retrieval and visualisation tools have been built, such as OMNIWeb, CDAWeb, PAPCO, SDDAS, KPVT, SPENVIS. A short description of these tools is given. The latest trend is to use data as input to expert systems that build instantaneous models of the space environment and its effects on systems (e.g. the SEDAT system under development at RAL).

The ever growing number of data bases available for scientific and other use necessitates the creation of search tools to select and locate data relevant for a given application. The increasing size of typical space weather data sets may render retrieval by Internet unacceptably slow, so that fast and flexible distribution systems on mass storage media will remain vital.

1. INTRODUCTION

Modern day space science is becoming more and more dependent on the vast quantities of data that are collected continuously from an ever increasing number of spacecraft and ground based measuring stations. Providing the international community with access to these data is stretching data archives and distribution systems to the limits of current day technologies. In general, data users expect the following from a data base provider service:

- performance: data requests have to be serviced with minimal delays, which is especially demanding for internet based services;
- reliability: continuous access has to be guaranteed;
- standardised access capabilities that return data in well-defined formats and do not depend on the data type or storage facility;
- adaptability: the data base environment must be flexible so that new data and new data types can be accomodated without major modifications to the query interface;
- validation: the data distributed to the user community is usually preprocessed. When new processing algorithms are applied to archived data, users have to be provided with the most recent version.
- documentation: availability of detailed descriptions of data processing and storage.

The current situation is that different data providers implement these requirements in very different ways, so that standardised access to data is still not fully achieved.

Data distribution can be divided into two broad categories: distribution by means of mass storage media such as CD-ROMs and DATs, and distribution via the Internet. Both methods have their advantages and drawbacks. Distribution in the form of hardcopies is reliable but not always fast, and requires storage of many copies of the data at different locations. Production of CDs usually is expensive for the typically small number of copies required, and automated storage devices are not within the reach of small institutes. Distribution via the Internet is faster in the sense that there are no production or mailing delays, but the bandwidth of the Internet is not sufficient to transfer very large amounts of data with acceptable download rates.

The two main components of a data archive are the data format and the software tools that query and process the data. In Section 2, we present short descriptions of the data formats that are most often used in space physics. This list is far from exhaustive, which is also true for the list of access and visualisation tools in Section 3.

Once data are archived, users have to be made aware of their availability. In our experience, it is often far from straightforward to locate a data set, even when one knows of its existence. In Section 4, we list some of the major data centres that are accessible through the Internet. There is an urgent need for general, flexible search tools to facilitate the search for data.

2. SCIENTIFIC DATA FORMATS

A variety of common data formats are currently used for the storage of scientific data. A comprehensive list is available at `http://www.cv.nrao.edu/fits/traffic/scidataformats/faq.html`. Most of these formats are included in the Open Information Interchange (OII) Standards and Specifications List. The next sections give an overview of some of the formats used by the space physics community. Formats which are specific to other disciplines are not included. One should note that the word format here does not mean physical storage in the sense of bits and bytes, but a way of representing data. As the different formats presented were developed by different communities, it is clear that each was developed with that community in mind.

2.1. CDF

The Common Data Format (CDF, Ref. 1) consists of a library and toolkit for storing, manipulating, and accessing multi-dimensional data sets. The format allows for easy exchange between different platforms, and is widely used in the scientific community. It is developed by the National Space Science Data Center (NSSDC), and is used in different scientific programs such as ISTP, CLUSTER, and space weather programs. Also, large quantities of data using this data format are available on the internet (e.g. OMNIWeb, CDAWeb, COHOWeb).

In general, the contents of a CDF file can be divided in two catagories. The first is a series of records comprising a collection of variables consisting of scalars, vectors and n-dimensional arrays. The second is a set of attribute entries (metadata, data base dictionary) describing the CDF in global terms or specifically for a single variable. The CDF documentation for VMS is available from `ftp://nssdca.gsfc.nasa.gov/cdf/` and for other systems from `ftp://nssdc.gsfc.nasa.gov/pub/cdf/`.

2.2. FITS

FITS (Flexible Image Transport System, Ref. 2) is the standard data interchange and archival format of the worldwide astronomy community. Although its name suggests it is an image format, it was primarily designed to store scientific data sets consisting of multidimensional arrays (1-D spectra, 2-D images and 3-D data cubes) and 2-dimensional tables containing rows and columns of data. The NASA/Science Office of Standards and Technology (NOST) Standard and user guide are available from `ftp://nssdc.gsfc.nasa.gov/pub/fits`.

2.3. HDF

HDF (Hierarchical Data Format, Ref. 3) is a self-defining file format for transfer of various types of data between different machines, developed at the National Center for Supercomputing Applications (NCSA) at the University of Illinois. The data model is based on the hierarchical relationship and dependencies among data. The basic structure consists of an index with the tags of the objects in the file, pointers to the data associated with the tags, and the data itself. HDF Files are difficult to update as data records are physically stored in a contiguous fashion. Therefore, if a data record needs to be extended, it usually means that the entire file has to be rewritten. Also, HDF only supports host encoding and XDR. The HDF homepage is at `http://hdf.ncsa.uiuc.edu/`.

2.4. netCDF

netCDF (Network Common Data Format, Ref. 4) Is an interface for scientific data access which implements a machine-independent, self-describing, extendible file format. It was developed a few years after CDF by the National Center for Atmospheric Research (NCAR) and was based on the CDF conceptual model. When it was developed, it provided a number of additional features, but with the current versions, both are quite similar in most respects. The existing netCDF software reads and writes data in only the XDR data encoding, and the internal caching algorithm can not be modified by the user to improve the performance. The home page for netCDF is at `http://www.unidata.ucar.edu/packages/netcdf/`.

2.5. VICAR

VICAR (Video Image Communication and Retrieval) Is a collection of image processing programmes supported by the Multimission Image Processing Laboratory (MIPL) at the Jet Propulsion Laboratory (JPL), for use in manipulating and analysing spacecraft images. The official description of the VICAR image format is available at `http://www-mipl.jpl.nasa.gov/vicar/vic_file_fmt.html`.

2.6. PDS

In recent years, PDS (Planetary Data System, Ref. 5) has been responsible for archiving space mission data on CD-ROM, using its own self-describing data format, variously known as PDS or ODL (Object Description Language). It was used in some missions (e.g. Magellan, Galileo) as a pointer to detached VICAR-format imagery on the CD-ROM volumes. This simple, versatile, human-readable, self-describing, object-based data format is used to encode metadata for NASA space science missions and others, such as the DNA Data Archival and Retrieval Enhancement (DARE) project. The PDS Standards Reference Document can be found at `http://pds.jpl.nasa.gov/`.

2.7. SAIF

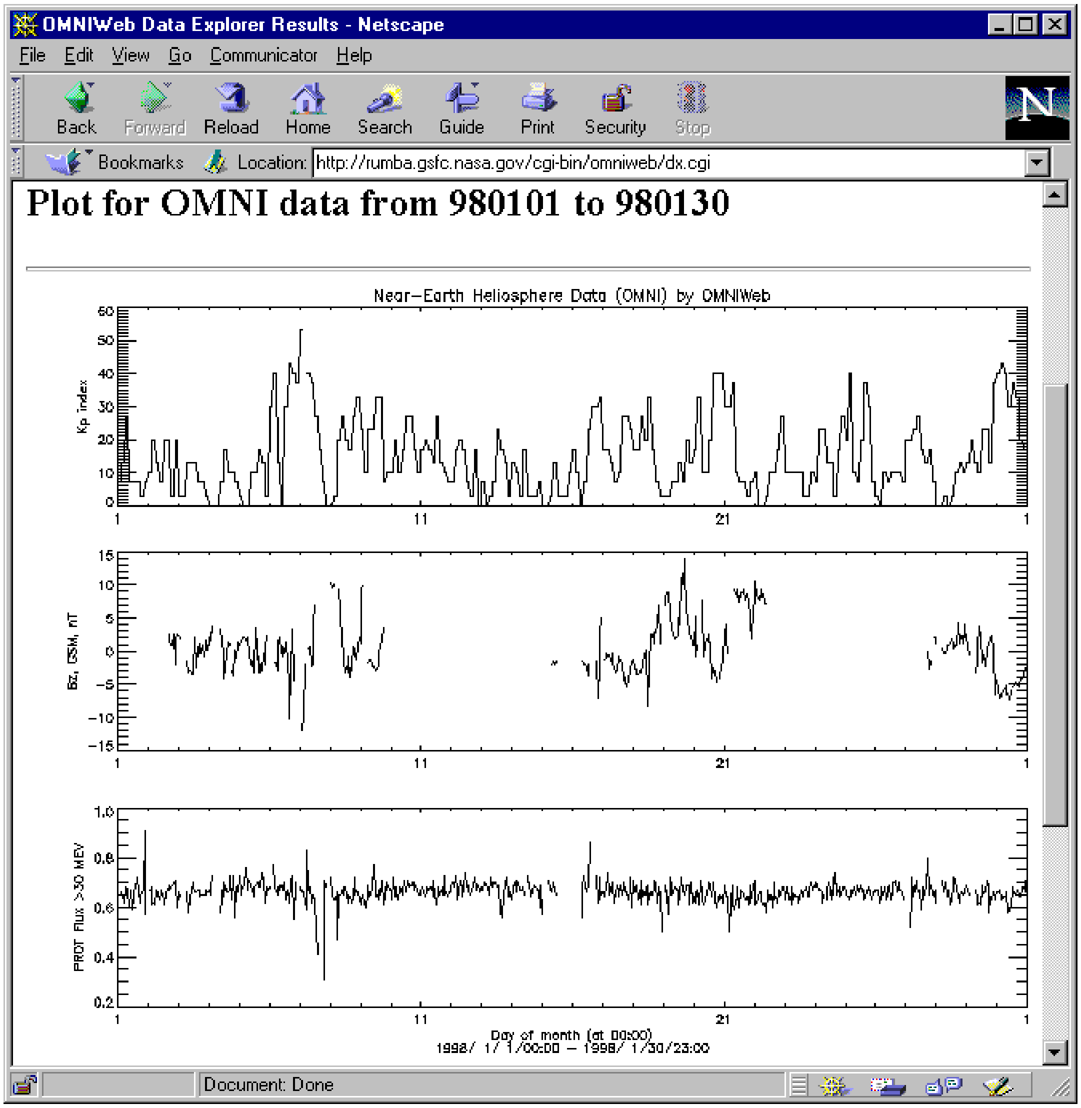
SAIF (Spatial Archive and Interchange Format) Is a Canadian standard for the exchange of geographic data. It uses an object oriented data model, and consists of definitions of the underlying building blocks, including tuples, sets, lists, enumerations, and primitives. The home page is available at `http://epoch.cs.berkeley.edu:8000/sequoia/schema/html/BigSur/saif/saifHome.html`.

2.8. SDTS

SDTS (Spatial Data Transfer Standard) Is a Federal standard (Federal Information Processing Standard (FIPS-173) for transfer of geologic and other spatial data. Documentation and examples are available from the U.S. Geological Survey (USGS) ftp-site at `ftp://sdts.er.usgs.gov/pub/sdts/`.

2.9. HDS

HDS (Hierarchical Data System) Is a freely available data base system that is particularly suited to the storage of large multi-dimensional arrays (with their ancillary data) where efficiency of access is a requirement. It is presently used in astronomy for storing (in particular)



Plot for OMNI data from 980101 to 980130

Document: Done

OMNIWeb Data Explorer Results - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Guide Print Security Stop
Bookmarks Location: http://rumba.gsfc.nasa.gov/cgi-bin/omniweb/dx.cgi
Near-Earth Heliosphere Data (OMNI) by OMNIWeb
Kp index
0 10 20 30 40 50 60
Bz, GSM, nT
-15 -10 -5 0 5 10 15
PROT Flux >30 MEV
0.2 0.4 0.6 0.8 1.0
1 11 21 1
Day of month (at 00:00)
1998/ 1/ 1/00:00 – 1998/ 1/30/23:00

Figure 1. *OMNIWeb sample output screen*

images, spectra and time series. It is developed by Rutherford Appleton Laboratories (RAL) and used in the Starlink project. Documentation and information on obtaining the source code, is available at `http://star-www.rl.ac.uk/cac/publicity/news_hds.html`.

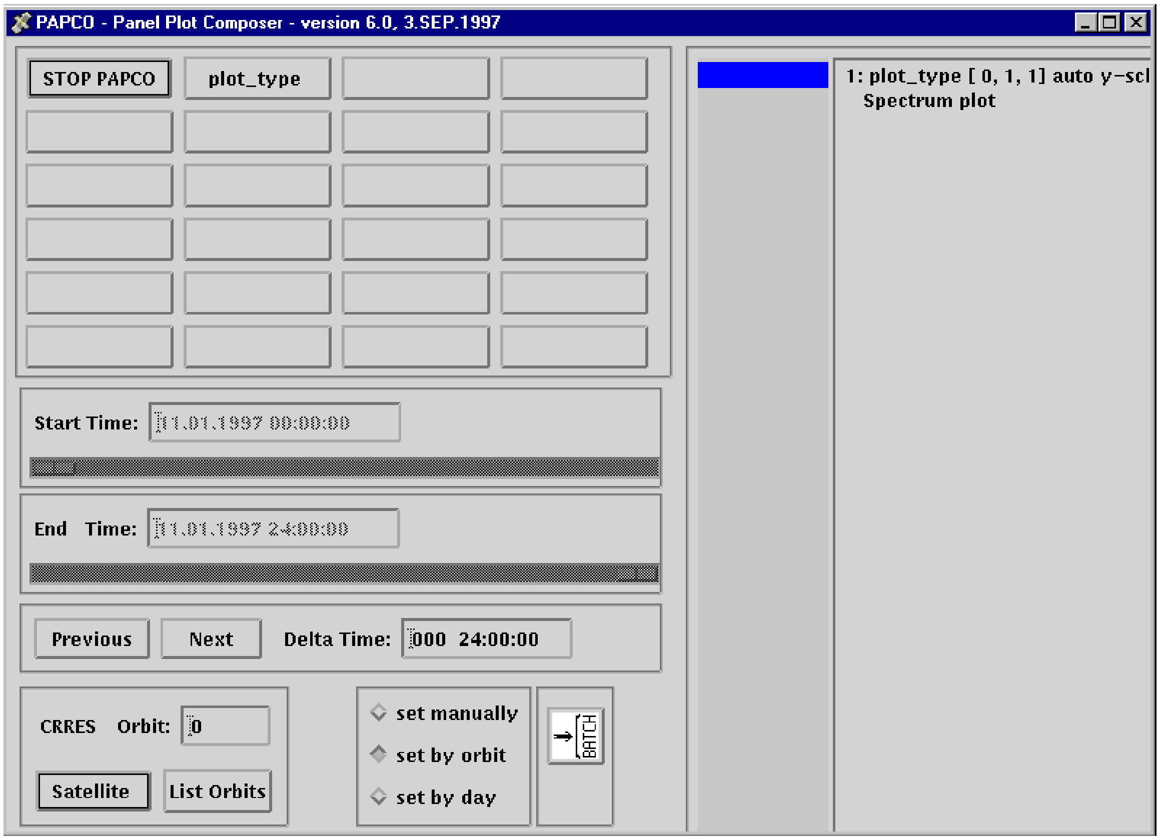
2.10. FFH/FFD

UCLA's Institute of Geophysics and Planetary Physics' (IGPP) Flat File System. is a two-dimensional data base system useful for storing data, such as time series, as rows of records each consisting of data fields in columns. The system is quite flexible because the data fields can contain most common computer data types.

2.11. SFDU

The Standard Formatted Data Unit (SFDU, Ref. 6) concept developed by the Consultative Committee for Space Data Systems (CCSDS) provides standardised techniques for the automated packaging and interpreting of data products. It puts no constraint on the format of the user data, and can thus accommodate standard formats developed by other organisations or user communities. It operates in a heterogeneous environment. The SFDU offers:

- a low overhead, internationally recognised data labelling scheme which permits self-identification of data objects;
- standard techniques for providing complete and unambiguous data descriptions;
- procedures for registration and administration of these data descriptions;
- techniques for packaging labelled data objects into larger data products;
- sufficient standardisation to allow the development of generic software to support the retrieval, access, parsing and presentation of SFDU data objects, while allowing those objects to have individual formats to satisfy particular application and user needs.



PAPCO - Panel Plot Composer - version 6.0, 3.SEP.1997

Start Time: 11.01.1997 00:00:00

End Time: 11.01.1997 24:00:00

Delta Time: 000 24:00:00

CRRES Orbit: 0

◇ set manually
◆ set by orbit
◇ set by day

1: plot_type [0, 1, 1] auto y-scl
Spectrum plot

Figure 2. *PAPCO Sample interface screen*

Information about SFDU is available at `ftp://nssdc.gsfc.nasa.gov/pub/ccsds/text/CCSDS-621.0-G-1.txt`.

2.12. IDFS

The Instrument Data File Set is a set of files written in a prescribed format which contains data, timing data and meta-data. IDFS development and evolution has been underway for the past decade, overlapping with the other storage and representation methods like CDF, HDF and FITS. The impetus behind the IDFS is the need to maintain certain metadata parameters with the data. The two key tasks supported by the IDFS are the conversion of telemetry values to engineering and scientific units and the registration of each data sample to a given point in time. IDFS covers a large quantity of calibration and timing factors, and can be exceedingly detailed, as many spacecraft have instruments that operate in nondeterministic ways due to data adaptive mode changing. IDFS requires at least three different files: a data file, a header file and the virtual instrument description file. Additional files to further parameterize the creation of displays or to perform user-specified transformations before plotting or displaying are also included. The format was developed by the Southwest Research Institute (SwRI) and is implemented in the Southwest Data Display and Analysis System (SDDAS). Information about IDFS can be found at `http://pemrac.space.swri.edu/spds/userguide/data.html`

3. VISUALISATION AND ANALYSIS TOOLS

A number of visualisation and analysis tools have been built around the data formats described in Section 2. In this section, we briefly describe some of the tools that are used most.

3.1. CDAWlib

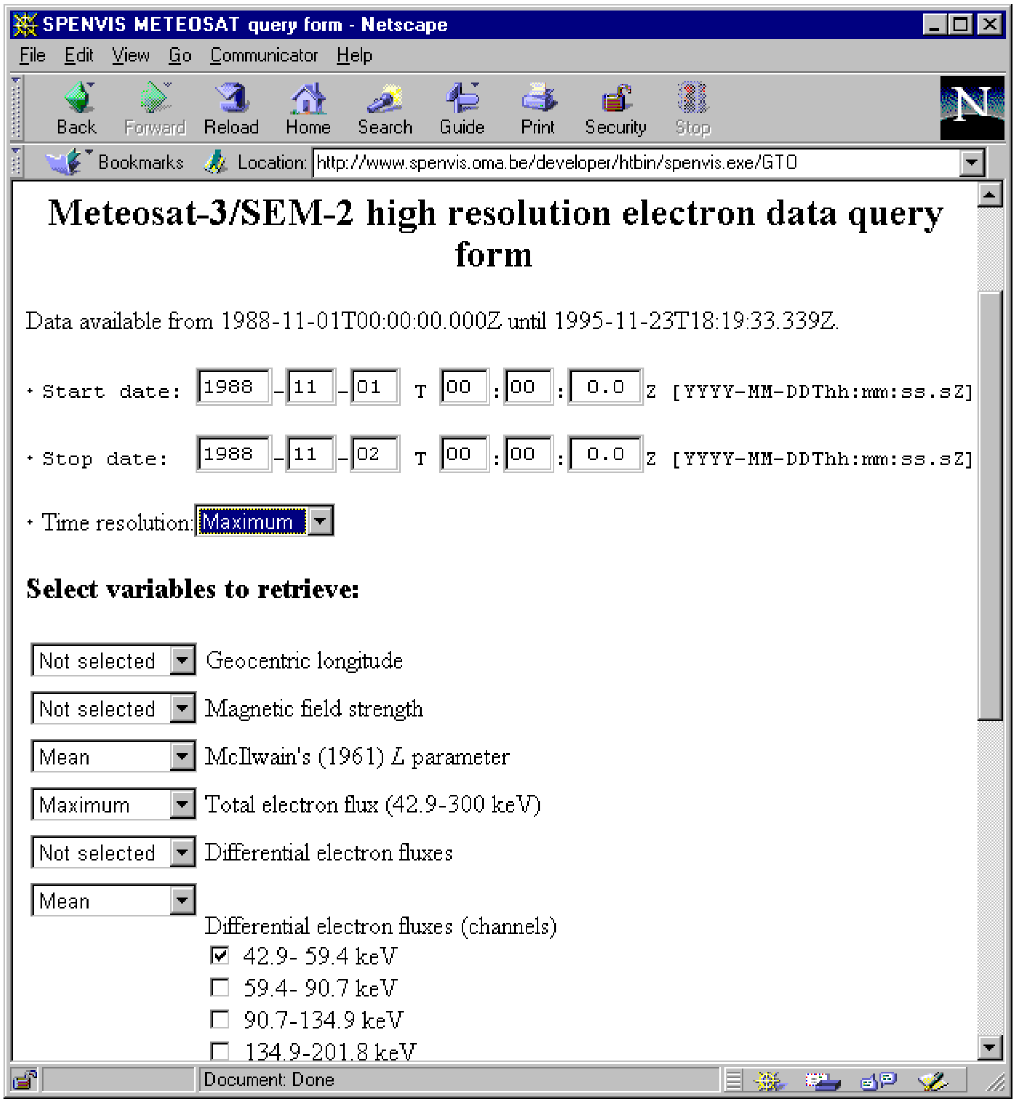
The Coordinated Data Analysis Web library (CDAWlib) consists of a library of Interactive Data Language (IDL) routines, using CDF for the storage of the databases.It uses the ISTP guidelines as logical format for the CDF files, with minor modifications. The version which is available on the WWW (`http://nssdc.gsfc.nasa.gov/space/spdf/CDAWlib.html`) allows only plots as a function of time.

CDAWLib Is used in CDAWEB (`http://cdaweb.gsfc.nasa.gov/cdaweb/sp_phys/index.html`).

3.2. KPVT

The Key Parameter Visualization Tool (KPVT), developed at the ISTP Science Planning and Operations Facility (SPOF), is a generic software package to visualize the key parameter data produced from ISTP missions, interactively and simultaneously. The tool is designed to facilitate correlative displays of ISTP data, and thus the selection of candidate events, data quality control.

The software, written in IDL, includes a graphical user



Location: http://www.spenvis.oma.be/developer/htbin/spenvis.exe/GTO

Meteosat-3/SEM-2 high resolution electron data query form

Data available from 1988-11-01T00:00:00.000Z until 1995-11-23T18:19:33.339Z.

- Start date: 1988 - 11 - 01 T 00 : 00 : 0.0 Z [YYYY-MM-DDThh:mm:ss.sZ]
- Stop date: 1988 - 11 - 02 T 00 : 00 : 0.0 Z [YYYY-MM-DDThh:mm:ss.sZ]
- Time resolution: Maximum

Select variables to retrieve:

Not selected Geocentric longitude

Not selected Magnetic field strength

Mean McIlwain's (1961) L parameter

Maximum Total electron flux (42.9-300 keV)

Not selected Differential electron fluxes

Mean Differential electron fluxes (channels)

☑ 42.9- 59.4 keV
☐ 59.4- 90.7 keV
☐ 90.7-134.9 keV
☐ 134.9-201.8 keV

Figure 3. *Sample screen of the SPENVIS data base interface*

interface, and runs on many platforms including multiple UNIX workstations, OpenVMS, Windows 95, and Macintosh. The full package of the tool, including the source code, installation instructions, and other material, can be obtained via anonymous FTP (`ftp://istp1.gsfc.nasa.gov/tools/kp_plot/`) or from the ISTP website (`http://www-istp.gsfc.nasa.gov/`).

A major drawback of KPVT is that it is based on X-windows widgets only, and not adapted for Web applications.

3.3. OMNIWeb

OMNIWeb Is a WWW-based data retrieval and analysis interface to NSSDC's OMNI data which consist of 1-hour-resolution near-Earth solar wind magnetic field and plasma data, energetic proton fluxes (1–60 MeV), and geomagnetic and solar activity indices. It allows the user to select a subset from the available OMNI data to view or retrieve. It also provides a graphical browsing capability to analyse and preview the data as time series plots. This browsing feature was designed to assist the user in following trends in the data and discovering areas of interest.

OMNIWeb Is available at `http://nssdc.gsfc.nasa.gov/omniweb/`. Figure 1 shows a sample output screen for OMNIWeb.

3.4. PAPCO

PAPCO Is a plotting software developed at the Max Planck Institut für Aeronomie. The aim of this PAnel Plot COmposer is to allow the user to put together data from a great variety of sources, to fully exploit the aims of the ISTP program. The package is described at `http://shaper.bnsc.rl.ac.uk:8080/ccr/software/papco/papco.html`

PAPCO is written in IDL and is modular in structure. It allows the user to select existing modules or construct new modules for a new dataset. The data for the different modules is not included in the distribution. The different modules use the datasets in their respective formats. This means that for each module, new reading and plotting routines have to be added to the package.

The current version of PAPCO is based on widgets. This is not suited for implementation on the WWW. Development for a WWW-driven PAPCO interface is underway, but not implemented yet.

3.5. SDDAS

The Southwest Data Display and Analysis System (SDDAS) is a set of X-window applications to display data from multiple data sources. It also provides access to a repository of data from the UARS, DE, CRRES, and TSS-1 programmes, among others. It makes use of the IDFS format. SDDAS Allows data in distributed archives, from many different satellites and other sources to be displayed and analysed using a diverse set of graphical applications. The SDDAS home page is `pemrac.space.swri.edu/spds/`.

3.6. The SPENVIS interface

We have developed a WWW data base visualisation and retrieval interface for ESA's SPace ENVironment Information System (SPENVIS), which is described elsewhere in this volume (Ref. 7) and can be accessed at `http://www.spenvis.oma.be`. The data sets to be implemented in SPENVIS have been converted into CDF format following the IACG/ISTP and CLUSTER guidelines. The interface uses a CGI script to interpret HTML form parameters and passes them on to an IDL programme that queries the data bases, produces output files and plots, and generates HTML code. A sample screen of the interface is shown in Figure 3. Ref. 7 contains a sample plot panel.

4. DATA CENTRES

Satellite and ground based data are available at a very large number of sites. It is not always straightforward to locate a given data set, even when one is aware of its existence and usefulness. Table 1 contains the URLs of a few data archives that store data from very different sources. The respective HTML pages contain links to many other related sites. As far as we know, there is no official site with complete listings of data servers and their contents.

There is an obvious need for powerful search tools that allow user input describing the type of data wanted. The search tools now in use generally are organised as a function of satellite mission, instrument, or research programme (as an example, Figure 4 represents the SPyCAT mission selection screen). However, more general queries for data are not possible. Locating data would become much easier if the international community could agree upon and implement a policy of data base indexing and cross-referencing.

5. FUTURE NEEDS

Based on our own experiences, we feel that new or continued developments are needed in the following areas:

- The data rates foreseen for future and even ongoing missions are daunting: several Gb per day will be no exception. Distribution of these data, especially in real time, will require vastly enhanced Internet bandwidths on the one hand and faster and more capacious mass storage devices on the other.
- To our knowledge, there are no mechanisms in place to guarantee that data downloaded by or sent to a user are up to date, i.e. that the data in the archive has not been modified. Protocols to rule out the co-existence of different versions of the same data should be developed and enforced.
- A further standardisation of data formats and access tools is necessary. Coordinattion efforts such as those of the IACG should be encouraged and extended. Also, new industry standards such as CORBA should be investigated.
- More flexible search tools than those currently operational are urgently needed. This implies in turn that all data archives are linked in some way, even if this is limited to their keywords and contents descriptions.

ACKNOWLEDGEMENTS

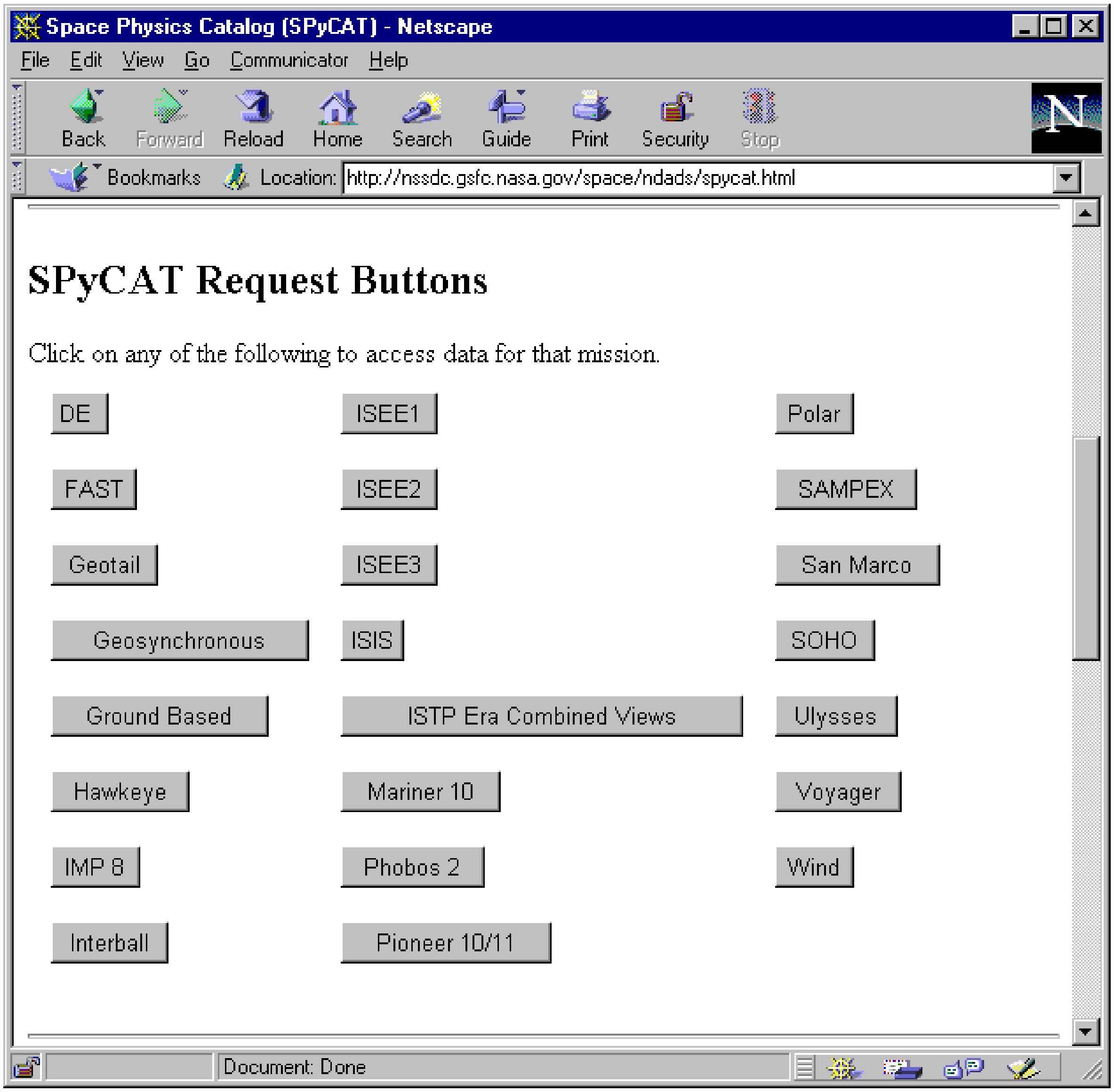
This work was funded by ESA/ESTEC/TOS-EMA General Support Technologies Programme Contract No. 11711-WO1.

REFERENCES

1. 1996, *CDF User's Guide*, Version 2.6, NSSDC, Greenbelt, Maryland.
2. Wells D C, Greisen E W & Harten R H 1981, FITS: A Flexible Image Transport System, *Astron. Astrophys. Suppl.*, 44, 363–370.
3. 1998, HDF 5 Reference Manual, NCSA
4. Rew R K & Davis G P 1990, NetCDF: An Interface for Scientific Data Access, *IEEE Computer Graphics and Applications*, 10, 4, 76–82.
5. 1995, *PDS Standards Reference*, Version 3.2, Jet Propulsion Laboratory, California Institute of Technology, Pasadena, California.
6. 1992, *Standard Formatted Data Units, A Tutorial*, CCSDS 621.0-G-1, Green Book.
7. Heynderickx D, Quaghebeur B, Fontaine B, Glover A, Carey W C & Daly E J 1998, New Features of ESA's SPace ENVironment Information System (SPENVIS), *this volume*.

Table 1. *Some data centres and services, with their WWW address*

Name	URL
World Data Centers	`http://www.ngdc.noaa.gov/wdc/wdcmain.html`
NASA Master Directory	`http://nssdc.gsfc.nasa.gov/nmd/nmd.html`
Space Physics Data System (SPDS)	`http://spds.nasa.gov/`
Space Physics Catalog (SPyCAT) for the ISTP ERA Data	`http://nssdc.gsfc.nasa.gov/space/ndads/istp.html`
Magnetospheric Yellow Pages	`http://nssdc.gsfc.nasa.gov/spdf/yellow-pages/`
ESA Archive for Ulysses Data	`http://helio.estec.esa.nl/ulysses/archive/`
Canadian Open Network for the Open Program Unified Study (CANOPUS)	`http://www.dan.sp-agency.ca/`
D_{st} Index	`http://swdcdb.kugi.kyoto-u.ac.jp/dstdir/`
Ionospheric phenomena	`http://www.wdcb.rssi.ru/WDCB/cat2.html`
Current Solar Forecast	`http://www.sel.noaa.gov/forecast.html`



Space Physics Catalog (SPyCAT) - Netscape

Location: http://nssdc.gsfc.nasa.gov/space/ndads/spycat.html

SPyCAT Request Buttons

Click on any of the following to access data for that mission.

Document: Done

Figure 4. *SpyCAT Request screen*