C D P P

Centre de Données de la Physique des Plasmas

&

GRID activity

C.C. Harvey

CNRS/CESR, Toulouse

What is the CDPP ?

The CDPP is a space physics data centre, which has three objectives :

- the long-term archiving of space physics data
- the provision of value added services
- facilitate reciprocal access to similar services elsewhere.

It is a French collaboration between CNRS and CNES, with :

- CNRS responsible for scientific aspects, such as the User Requirements, the identification and recovery of data and information to be archived, and the definition of the value added services.
- CNES responsible for technical aspects, such as the software system, formats, and standards for archiving and exchange.

CDPP development started in 1996, and the service was fully opened in October 1999. Since then the CDPP has pursued regular ingestion of data and addition of new services.

http://cdpp.cesr.fr **cdpp@cesr.fr**

What is meant by data archiving ?

Archiving consists of :

- The preservation of
  - the data and ancillary information in order that scientifically meaningful parameters, in scientific units and ready for use, can be recovered over an undetermined number of years.
  - Documentation describing how these parameters were measured, so that the user can formulate his own opinion as to the quality of the data, and for what purposes it can be used. For example
    - » electron density can be determined by a particle counter, a Langmuir probe, an active wave experiment, or from the observation of the naturally occurring wave spectrum
- A means to consult the documentation, and select then recover the data.

In other words, the archive must preserve the data and **ALL** the information required to interpret the data scientifically.

What are analysis tools ?

- An analysis tool is a piece of software which performs a well-defined specific <u>scientific</u> computational task, such as :
  - averaging of data
  - rotation of coordinates
  - variance analysis
  - Fourier transformation
  - other tools to analyse the data (not at CDPP).
- It may require more than one data product on input, for example,
  - for cross-correlation
  - for calculation of the β of a plasma
- It should be generic, that is, applicable to any physical parameter of the same type (scalar, vector, tensor).
- The output may be digital (*i.e.*, another data product) or graphical.

Data centre tools

In addition, a data centre must provide "data centre tools" :

- Tools to help locate the documentation and/or data
- Tools to visualise the data :
  - prior to delivery
  - from one or several archived data sets
- Tools to recover segments of the data with :
  - decommutation of the parameters required
  - transformation of coordinate system, format, or time synchronisation
- The possibility to **update** data or its documentation (including caveats)

A data centre may also provide :

- Tools to perform operations on the data before delivery
  - resampling, rotation of coordinates, etc.
- Software packages, for example :
  - models of the Earth's magnetic field
- Other services, such as calculation of satellite rendez-vous.

Interoperability

No one centre can archive data or provide all services for the entire world !

1) “All data” or “services” for what discipline ? However disciplines are defined, there will always be borderline cases.

2) Even when disciplines are defined, it rapidly becomes clear that the quantity of data to be archived far exceeds the ability of any one centre to handle it.

3) Likewise, no one agency will wish to take responsibility for funding one massive centralised data centre.

4) Data centres also provide services to the end users, which facilitate the exploitation of the archived data. The development and continuing evolution of these services require the participation of an active local scientific research community.

For all these reasons data centres will become increasingly dispersed across the globe. Interoperability is required.

This is where the GRID may be used.

What is Interoperability ?

- The science user requires information, data, and services.
- His local interface should appear as the access to a single coherent global system, which in reality it consists of
  - many geographically dispersed centres,
  - which store data from different missions or different experiments,
  - possibly in different formats, and
  - offer different value-added services.
- The objective of "Interoperability" is to allow the user, wherever he is situated, to find via a graphical interface with which he is familiar:
  - the information he wants,
  - the data he needs, and
  - the tools to exploit it,
  - all in a form which he can readily use.
- Thus interoperability creates a "Global Virtual Observatory"

Data & service descriptions

Today, to find some specific data or service, the scientist must visit several data centres, each with its own Web interface.

This is wasteful of time ; in principle, a single query should be enough.

To advance, it must be noted that :

- To search for something, we must be able to describe it !
- For across-the web searching, this description must be machine interpretable.
- Keywords are not enough, they must be structured. This is the “new” discipline of ontology.
- All data centres must provide and maintain the information which describes each of their data sets.

Then it is, in principle, possible to search for data, as follows.

Data Searching

A scientist initiates a search request :

- The search information is converted to a standard form which is sent to all data centres <u>of that discipline.</u>
- Each data centre uses its own search engine to formulate a response which is returned in a standard format
- The scientists "local" centre concatenates all the replies,
  - taking account of the ability of each centre to use the given search criteria (otherwise the least discriminating centre will score the highest number of "hits")

  and presents the results on a single Web page.

Note that :

- most Interoperability software will be common to all data centres ;
- each centre remains free to develop its own system of data archiving and data restitution.
  - This is an essential requirement, because normally a data centre will "exist" before wishing to participate in interoperability.

Data Recovery

- Having found where the required data exists, it remains :
  - to select the exact data segment required, then
  - to recover it.

- Interoperability should eventually make both data selection and recovery possible without the need to access the remote data centre where the data actually resides.

- It would be useful for data centres to offer a selection of standard formats for data delivery. Suggestions :
  - plain ASCII (which seems to be the most popular)
  - CDF
  - netCDF
  - other

Data Tools

- Our heritage is catastrophic : a totally fragmented system.
- Each major laboratory has its own analysis tools.
- The investment in development is such that most labs are highly unwilling to change in the near future.
- A plausible medium-term solution could be the adoption of a standard format for the exchange of data and metadata :
  - to permit data to be shipped from one laboratory to another
  - with both the data and metadata
  - in machine assimilable form (to plug directly into applications)
  - All self-respecting data tools (and archives) should be able to both input and output data in this format
  - Once widely used, commercial software would start to use it

The big problem - who specifies this format ?

Coordination in Europe

- Individual data centres are already archiving data. They need to know that their effort :
  - is not being duplicated elsewhere,
  - is compatible with on-going developments elsewhere,
  - has the support of the international community.
- The earlier data archiving is prepared the better, because early preparation allows :
  - maximum reusability of code → reduced testing, increased reliability
  - access to the engineers who really know the experiment
  - a smooth transition from the science operations to the archival mode of science analysis
- For both reasons (compatibility and forward planning) data centre activity requires coordination.
- These remarks apply to all disciplines (not only to space plasmas).

Conclusion

Data centre activity requires NEITHER intensive processing
NOR Gbytes/sec of network bandwidth.

It DOES require :

- TECHNICAL COORDINATION
  - intelligent and machine intelligible exchange of information over the network ;
  - standard formats for exchange of data, metadata and service information
- MANAGEMENT COORDINATION
  - coordination of the archiving activity, to decide who archives what, and where.
  - forward planning, so that archiving can be prepared in parallel with the planning of the science operations.

This coordination is required at global, continental and national level.

Can such coordination in the realm of space research be considered a GRID activity ?

End of the presentation

Thanks