Centre de Données de la Physique des Plasmas

Les dictionnaires de métadonnées en Physique Spatiale

Christopher C. Harvey

Centre de Données de la Physique des Plasmas

http://cdpp.cesr.fr

cdpp@cesr.fr

Le CDPP

Le CDPP est une Collaboration CNRS/CNES :

- Responsabilités **CNES** :
  - Réalisation du SIPAD (Système Informatique de Préservation et d'Accès aux Données)
  - Maintien en fonctionnement et gestion du SIPAD (insertion données, autorisation d'accès, suivi des commandes, évolutions logicielles, *etc.*)
  - La préservation des fichiers.
- Responsabilités **CNRS/INSU** :
  - Choix des données à archiver
  - Facilité d'utilisation de l'interface graphique de l'utilisateur
  - Services à Valeur Ajoutée
  - Validation des demandes d'inscriptions et des droits d'accès
  - Pages de vulgarisation (le public est bien intéressé)
- Responsible Technique : Claude Huc, CNES/CST
- Responsable Scientifique : Cristopher Harvey, CNRS/CESR

Les Missions Spatiales

Les missions scientifiques spatiales peuvent être classées en deux catégories bien distinctes :

- Missions de type « Observatoire »
- Missions du type « Principal Investigator »

- Il y a également des missions du type
  - communications,
  - GPS (localisation),
  - et d'autres

  dont les informations principales ne sont pas vraiment « spatiales ». Néanmoins on garde des paramètres de surveillance, qui tombent généralement dans la catégorie « Principal Investigator ».

Les Missions Spatiales du type « Observatoire »

Missions de ce type font généralement de la télédétection :

- Les observations sont programmées à l'avance :
  - Observation d'une galaxie, une étoile, ou un autre type d'objet céleste ;
  - Observation d'une région de la surface de la Terre.
- Les observations peuvent être programmées par un « Guest Observer »
- Il y a concertation entre les observations faites par les différents instruments (qui visent tous la même direction).
- En conséquence :
  - les résultats peuvent être caractérisées par phénomène observée, direction d'observation, ou les deux (un peu comme pour un observatoire au sol).

Les Missions Spatiales du type « Principal Investigator »

Missions de ce type font des observations du milieu ambiant autour de la sonde spatiale :

- Plusieurs instruments sur une seule sonde mesurent et enregistrent le long du trajectoire les différents paramètres physiques, telles que :
  - **le champ magnétique,**
  - **la densité du plasma, sa composition,**
  - **le spectre d'ondes électomagnètiques, *etc.* ;**
- Souvent ces observations se font sans grande concertation entre les responsables des différents instruments, parce que
- A priori on ne sait pas ce qu'la sonde va rencontrer, puisque les conditions sont très fluctuantes, et donc
- On essaie d'identifier ce qu'on a observé (onde de choc, discontinuité magnétique, turbulence non-linéaire, *etc.*) soit
  - **à partir des données reçues, soit**
  - **par défaut, par rapport à un modèle statistique.**

Les Données Spatiales

On voit immédiatement les conséquences pour l'archivage des données issues de ces deux types d'expérience :

Il y a deux types de données spatiales

- Celles issues des missions de type « Observatoire » :
  - Données « orientées objet », ce qui facilite leur description.
  - Données souvent en forme d'image → formats standards (notamment le FITS, Flexible Image Transport System).
- Celles issues des missions de type « PI » :
  - Données organisées en séquences temporelles très longues (années) sans possibilité de savoir facilement ce qu'on observe à chaque moment.
  - Données avec formats très différents, même quand il s'agit du même paramètre issu d'une expérience différente.
  - Aucune standardisation → aucun outil générique répandu.

Définitions appliquables aux données « PI »

- **Jeu de Données (Dataset)** : une séquence de données de la même nature, issues du même instrument et produites par les mêmes logiciels. Un jeu de données contient généralement des centaines, voire des milliers, de fichiers.
- **Granul**e : l'élément le plus petit qui peut être identifié dans un dataset.
- **Données auxiliaires** : les données essentielles pour interpréter les données scientifiques : orbite, attitude, éclipses, température, alimentation électrique, *etc.*, en forme de jeu de données.
- **Orbite** : Le trajectoire suivi pas le satellite ou sonde. Egalement utilisée pour le jeux de données constitué par les coordonnées successives de la position du satellite en fonction du temps (comme tous les autre autres jeux de données « PI »).

Noter la différence par rapport aux expériences type « Observatoire », où un « jeux de données » identifie souvent toutes les données relèvent d'une observation particulière.

La Description des Données « PI » (1/2)

On peut décrire les données « PI » à trois niveaux :

- Jeu de données
  - Tout ce qu'on puisse dire a propos des données sans ouvrir des fichiers d'orbite ou examiner préalablement les données.
  - C'est un bon début, mais le scientifique à la recherche de données souhaiterait mieux.
- Granule (le plus petit élément d'un dataset qu'on peut définir)
  - On ouvre, par exemple, les fichiers d'orbite pour savoir où le sonde était réellement à un temps donnée.
  - On peut créer (à partir d'une étude des données elles-mêmes) des catalogues d'événements pour identifier les différentes régions de l'espace et leurs frontières (ondes de choc, *etc.*)
- Intermédiaire (entre les deux autres)
  - On peut utiliser les paramètres qui décrivent l'orbite pour savoir si, par exemple, c'est la bonne saison pour trouver les données souhaitées.

La Description des Données « PI » (2/2)

- Le seule niveau qui peut toujours être décrit est le jeu de données.
- On essaie de le décrire avec une maximum de précision. Pour ceci on utilise :
  - Data Modèle : une hiérarchisation des concepts qui caractérisent le jeu de données. Cette hiérarchisation est nécessaire pour :
    - » **Permettre de mieux composer les requêtes lors de la recherche des jeux de données.**
    - » **Éviter la redondance des informations quand on archivage plusieurs jeux de données en provenance d'un même satellite, ou institut ;**
  - Data Dictionary : la liste de mots qui caractérisent cette hiérarchie. Une liste est nécessaire pour pouvoir chercher par mot clé.
- Comme exemple, j'utiliserai le Metadata Dictionary développé pour Cluster Active Archive.

La Description du Dataset : le CAA Metadata Dictionary

- Cluster Active Archive (CAA) is a project of the European Space Agency, aimed at facilitating and encouraging the analysis of data from the Cluster mission (four identical spacecraft orbiting the Earth in a tetrahedral configuration).
- The CAA Project required a data dictionary to be ready for the negotiation of Interface Control Documents between the Project and each of the 11 Cluster Principal Investigators.
- Issue 1.0 of the Metadata Dictionary is dated September 21, 2004. This version is available at the URL

  `http://www.cesr.fr/~harvey/DataD_0_10.pdf`
- Completion and revision are in progress.
- This version describes data at the dataset level.
- This metadata dictionary will also be used for the next generation of CDPP.

The Data Model (1/3)

The data model is hierarchical : numerical data requires 6 levels for its description :

1. Mission : for CAA, the only possibility is Cluster
2. Observatory / Experiment
   - » These two are placed at the same level for Cluster
   - » There are 4 observatories (spacecraft) and 11 experiments
3. Instrument : the physical instruments belongs to both an observatory and an experiment, for example CIS1, FGM3, …
4. Dataset : each instrument will produce one or more datasets
5. Parameter : each dataset will contain one of more parameters
6. File : data is preserved in files. Information concerning these files is for CAA internal use, the end user never needs it.

The Data Model (2/3)

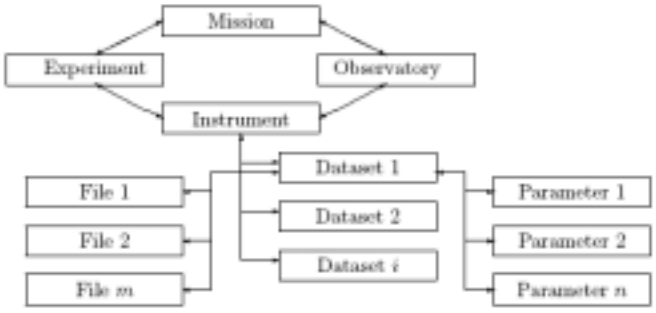

Figure 1: The CAA hierarchy of concepts. The arrows indicate the ascending and descending links. The n parameters and m data files are indicated only for Dataset 1 ; a similar hierarchy will exist for all i datasets of each instrument.

The Data Model (3/3)

The relationship between the various levels is defined by ascending keywords : for example, the dataset level has the predicate "Experiment = *value* " where *value* defines the experiment from which the dataset has been derived.

Each level inherits the properties of the higher levels upon which it depends.

In addition to numerical data, the following types of data product will also be archived :

- Graphical : preplotted graphics for browsing,
- Events : these are characterised by two times : start & stop,
- Documents : describing the various holdings,
- Software : related to the Cluster mission.

They are also integrated into the metadata hierarchy.

Description of the Parameter (1/4)

Ce qui intéresse le chercheur pour faire avancer sa recherche est le parametre mesuré : densité des électons, champ magnétique, … (plus le lieu et la date de la mesure, mais ça c'est une autre histoire !)

The description of the parameter is flat. It consists of :

- Entity.
  - What is observed : electron, proton, neutral, magnetic field, electric field, …
- Property
  - The property which is observed : number_density, flux_density, velocity, temperature, vector, magnitude, …
  - Not all properties can be associated with all entities, but – so what, does this matter ?
- Fluctuations
  - Often it is not the actual property which is measured, but some statistical property of its fluctuations, such as variance, spectrum, cross-correlation, cross-spectrum, covariance, polarisation, …

Description of the Parameter (2/4)

Vectors and Tensor Quantities

- Physical observables may be scalars, vectors or tensors, *i.e.*, tensors of order 0, 1, 2, or higher.
  - Tensor order is a useful property to identify, because all tensors of the same order transform in the same way.
- Sometimes a vector may be have one component missing, *e.g.*, along the spin axis ; missing elements are identified.
  - Some standard operations may still be possible, such as rotation about the direction of the missing component.
- Tools exist to extract a single compenent of any vector or tensor quantity.

Description of the Parameter (3/4)

- The coordinate system used to represent any vector or tensor quantity must be described.
  - The following coordinate systems are allowed for Cluster : GSE, GSM, SC (SpaceCraft), SR2 (despun Spin Reference), and ISR2 (Inverted SR2, which is close to GSE for Cluster).
- Vectors may be represented in Cartesian or polar coordinates (spherical or cylindrical).

Description of the Parameter (4/4)

Other information is required for each parameter :

- Processing level
  - Raw, uncalibrated, calibrated, derived
- SI_Conversion (for calibrated data) ;
  - the factor by which the numerical value must be multiplied to express the quantity in its standard SI unit which, moreover, is identified. Example : 1.0e–9>T for the magnetic field in nT.
- Value_type : the type of information in the field
  - Float, double, integer, char, ISO-time, time-span
- Displaytype : the recommended method of display :
  - Time series, spectrogram, stack_plot

Compound parameters

- Some parameters, *e.g.*, the plasma β, are derived from data from more than one instrument.
- Such parameters have the predicate

  "compound = *list* "

  where the *list* of parameter IDs identifies the parameters used. (Each parameter has a unique ID.)
- Thus all the relevant information can be obtained, at least for parameters derived from CAA datasets.
- More study required for parameters derived from data archived across multiple data centres !

Event Tables

La recherche d'événements (data granules) ne peut se faire que par un des méthodes suivants :

1. En ouvrant les fichiers des jeux de données pour chercher les données qui satisfont les conditions demandées. C'est un processus très long.
2. À partir des données orbitales et une modèle statistique de la magnétosphère (la région de l'espace autour de la Terre où les conditions sont dominées par le champ magnétique terrestre), identifier les temps probables de rencontre d'un événément « normal ».
3. En utilisant les tables d'événements. Les tables d'événements doivent être construits préalablement par un processus de « data mining ».

L'utilisation d'« tables d'événements » semble le plus satisfaisant, mais comporte néanmoins des difficultés sérieuses :

- La création de ces tables n'est pas facile, surtout parce que
- l'identification automatique du phénomène recherché n'est pas toujours claire (*e.g.*, chocs parallèles, discontinuités de contact, …)

Interoperability

- The science user requires information, data, and services.
- His local data system should appear as part of 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"

SPASE

- Space Physics Archive Search and Extract (SPASE) is an international consortium of space physics archiving organisations.
- Developing a data model and recommendations for implementation of that data model, which will
- Permit data centres to implement tools to allow their science users :
  - to find space physics data of interest, wherever it is,
  - intercompare the data found, and
  - retrieve selected datasets or portions of datasets.
- Search across multiple data centres through a single query
- With a homogeneous presentation of the results, using common terminology, and taking account of each centre's ability to discriminate when searching.

- SPASE will demonstrate the feasibility of the system.
- SPASE will not run the individual archives, which remain the responsibility of their parent organisations.

Principal Participants

- CNES/CNRS Plasma Physics Data Centre (CDPP)
- NASA/National Space Science Data Center
- Rutherford Appleton Laboratory
- Southwest Research Institute
- Planetary Data System - UCLA Plasma Physics Interactions Node
- Applied Physics Laboratory

Prochains Etapes

- Un « Observatoire Virtuel » est un ensemble de « dépots » de données (" repositories " en anglais) qui paraissent pour l'utilisateur comme une seule base de données.
- L'utilisateur a un accès uniforme à l'ensemble de données archivées.
- En 2005 la NASA va lancer un Appel d'Offres pour des « Observatoires Virtuels » pour le Système Solaire.
- On attend des propositions pour :
  - Virtual Solar Observatory
  - Virtual Heliospheric Observatory
  - Virtual Magnetospheric Observatory
  - Virtual ITM Observatory (Ionosphere, Thermosphere & Mesosphere)
- Ces observatoires du Systeme Solar doivent également avoir un certain niveau d'« interopérabilité » entre eux.
- Les differents groups américains se préparent activement.
- Il y aucune coordination en Europe.

Problèmes actuels

1) Nos collaborateurs SPASE aux USA sont financés par le programme LWS / RT&T de la NASA – mais pas nous.

2) En Europe la communauté de la physique de plasma est très fragmentée :
   - L'ESA (ou la NASA) finance et gère les plates-formes pendant la fabrication et les opérations) ;
   - Les agences spatiales nationales financent les expériences embarquées ;
   - D'autres instances emploient les chercheurs ;
   - Et maintenant l'Union Européenne finance des projets associés (EGSO, EuroPlanet, …).

3) Qui est propriétaire des données : sûrement le contribuable.
   Mais cela ne suffit pas pour arriver à les faire archiver !

4) Un centre doit connaître ce qui se passe ailleurs :
   - Qui archive quoi, pour s'assurer que rien n'est oublié ;
   - Quels nouveaux centres sont en gestation, pour éviter la divergence de standards ou de technologie, et ainsi faciliter l'interopérabilité.

Future Plans

- Start using the data model. This is essential, for two reasons :
  - to test the SPASE and CAA data models on real data ;
  - To compare the CAA and SPASE models ;
  - to have some metadata so as to start breadboarding an implementation of the SPASE recommendations, and use the CAA dictionary for the SIPAD-NG to be used for CDPP.
- Pour les données existantes, traduire le PVL utilisé par le SIPAD actuel en XML utilisé par le SIPAD-NG.
- Participer à la réalisation d'une implémentation préliminaire de SPASE.
- Espérer que l'organisation européenne devient plus simple,
- et qu'un centre européen, géographiquement distribué, de données plasma voit le jour (tout seul, ESA ne pourra pas tout faire).

Conclusion

- Several Metadata dictionaries for Space Plamsa Physics are under construction.
- Experience from Cluster shows that each data centre must have it own metadata dictionary, so as to master of its own internal affairs :
  - timetable for implementation ;
  - level of detail of the descriptions to be maintained ;
  - nature of the contents.
- Various data centres in the same discipline must be interoperable, to create a "Virtual Observatory"
- Virtual Observatories in related disciplines (*e.g.*, all those in Solar System Physics) must also have some level of interoperability.
- All interoperability depends upon existance of an adequate metadata dictionary. The principal difficulty is to obtain the support required from the scientists who know the data.

End of the presentation

Thanks

Présentations utilisées pour la préparation de celle-ci

- \CAA\Presentations\CAAMetadataDic.ppt
- \CDPP\Transpar\PNST_041019.ppt