**DGA / CELAR**

# ETUDE SUR LA STANDARDISATION DES FORMATS DE DONNEES

**CS Systèmes d'Information**

Nomenclature :
Edition : **01** Date : **14/04/00**
Révision : **00** Date :

Réf. : SCS/EAST/ET Date : 14/04/00

## LOGICIELS LIBRES ET SECURITE DE SYSTEMES INFORMATIQUES
## ETUDE SUR LA STANDARDISATION DES FORMATS DE DONNEES

| | | |
|---|---|---|
| **Rédigé par** :<br>Equipe EAST | le 14/04/2000 :<br>CS Systèmes d'Information | |
| **Validé par** :<br>O. RENAUX | le 14/04/2000 :<br>CS Systèmes d'Information | |
| **Pour application** :<br>O. RENAUX | le 14/04/2000 :<br>CS Systèmes d'Information | |

**Pièces jointes :**



CS SI
13, rue Villet – ZI du Palays
BP 4042 – 31029 Toulouse Cedex 4 - France

## DIFFUSION INTERNE

| | | **Observations** |
|---|---|---|
| DGA | F. RIOUX | 2 exemplaires |

## DIFFUSION EXTERNE

| | | **Observations** |
|---|---|---|
| CS SI | O. RENAUX | 1 exemplaire |

# BORDEREAU D'INDEXATION

| | |
|---|---|
| CONFIDENTIALITE : AD | MOTS CLES : ETUDE, STANDARDISATION, DONNEES, EAST |

TITRE : Etude sur la standardisation des formats de données

AUTEURS : Equipe EAST (C.PELIZZARRI, O.RENAUX)

RESUME : Ce document regroupe les résultats de l ’étude sur la standardisation des formats de données, effectuée dans le cadre plus général de l’étude sur les logiciels libres et la sécurité des systèmes informatiques.

SITUATION DU DOCUMENT : Ce document vit seul.

| | | | | |
|---|---|---|---|---|
| VOLUME : 1 | PAGE : 53 | PLANCHES : / | FIGURES : 1 | LANGUES : F |

CONTRAT : Marché n°99-42-257

SYSTEME HOTE : PC/WINDOWS NT/WINWORD 97/ETUDE.DOC

# MODIFICATION

| ETAT DOCUMENT | | | | PAGES REVISEES | |
|---|---|---|---|---|---|
| **ED.** | **REV.** | **DATE** | **REFERENCE ORIGINE (pour chaque édition)** | **ETAT PAGE *** | **NUMEROS DES PAGES** |
| 01 | 00 | 14/02/00 | | | Création du document |

* : I = Inséré S = Supprimé M = Modifié

# SOMMAIRE

# 1. GLOSSAIRE

| | |
|---|---|
| ASCII | American Standard Code for Information Interchange<br>Ce standard d'encodage des caractères est utilisé intensivement dans la transmission de données. |
| ANSI | American National Standards Institute<br>Ce groupe est membre de l'organisation américaine qui appartient à l'ISO. |
| API | Application Programming Interface |
| CCSDS | Consultative Committee for Space Data Systems |
| CDF | Common Data Format |
| DDL | Data Description Language |
| DDR | Data Description Record |
| DED | Data Entity Dictionary |
| DEDSL | Data Entity Dictionary Specification Language |
| EAST | Enhanced Ada Subset |
| FITS | Flexible Image Transport System |
| HDF | Hierarchical Data Format |
| HTML | Hypertext Markup Language (langage de description des pages WEB) |
| IHM | Interface Homme-Machine |
| ISO | International Standards Organization<br>L'ISO est une organisation supportée par l'industrie établissant des standards internationaux pour tout ce qui est utilisé dans l'échange de donnée, en établissant des spécifications. |
| OASIS | Outil d'Aide à la Structuration des Informations Spatiales |
| PVL | Parameter Value Language |
| RTF | Rich Text File (format d'échange de textes défini par Microsoft) |
| SGBD | Système de Gestion de Base de Données |
| SGML | (Standard Generalized Markup Language), norme ISO de structuration de document utilisant une grammaire particulière nommé DTD. |
| W3 ou WWW | World Wide Web |
| XML | Extensible Markup Language (langage intermédiaire entre SGML et HTML) |

## 2. DOCUMENTS DE REFERENCES

DR1 CSDS 644.0-B-1: The Data Description Language EAST Specification (CCSD0010). Blue Book. Issue 1. May 1997.
Recommandation adoptée en tant que norme ISO/FDIS 15889.

DR2 CCSDS Panel 2 Overview
CCSDS 600.0-G-0.2 Draft Green Book – Nov 1996

DR3 CCSDS 647.0-R-1: Data Entity Dictionary Specification Language.
Red Book. Issue 1. November 1996.

## 3. TERMINOLOGIE ET DEFINITION

Ce paragraphe définit les termes principaux employés dans le cadre de cette étude.

| | | |
|---|---|---|
| Dictionnaire | | Un dictionnaire est une collection de définitions sémantiques de différentes entités de données. Ces définitions sont composées de quelques attributs obligatoires et d'attributs optionnels. Un dictionnaire peut être élaboré pour un produit ou peut être associé à une discipline. Dans ce dernier cas il contient un ensemble prédéfini de définitions d'entités pouvant être utilisées par les concepteurs et les utilisateurs de données en tant que références. |
| Discriminant | | Un discriminant est un composant d'une structure dont la valeur influe sur le contenu de la structure. |
| Document | : | Objet textuel possédant le statut de « produit ». Par exemple une publication scientifique, un manuel technique. |
| Enuméré | | Ensemble contenant un nombre restreint de valeurs discrètes, où chaque valeur discrète est nommée et unique dans cet ensemble. |
| Identificateur | | Une séquence de caractères qui désigne quelque chose. |
| Internet | : | Réseau de communication international, basé sur Arpanet et utilisant le protocole de communication IP. Internet. |
| Interopérabilité | | Possibilité d'échanger (des dictionnaires). |
| Intranet | | Réseau interne à une entreprise utilisant les mêmes technologies que l'Internet. |
| Métadonnées | : | Données décrivant des données. Ce terme peut être utilisé pour désigner un fichier de données auxiliaires associé au fichier de données à traiter (données de positionnement d'un satellite par exemple) ou bien pour désigner un fichier décrivant la syntaxe ou la sémantique des données. Dans le premier cas les métadonnées sont elles aussi un fichier de données, dans le deuxième cas, il peut s'agir d'un descriptif écrit en EAST. |
| Produit de données | | Une collection de un ou plusieurs fichiers de données, « packagés » pour une application |
| Sémantique | | Information qui définit la signification plutôt que la représentation physique des données. La sémantique couvre un large spectre de domaines, des informations simples du type «unité » aux informations plus complexes telles que les relations entre une entité et une autre. |
| Sous type | | Un sous type est un type associé à une contrainte, contraignant les valeurs du type à satisfaire à certaines conditions. Les valeurs d'un sous-type sont un sous-ensemble des valeurs du type. |
| Structure | | Une structure est un type composé contenant zéro ou plusieurs composants nommés, pouvant avoir des types différents. |
| Syntaxe | | Information définissant la représentation physique des données. Elle inclut la composition structurelle des champs dans la donnée, sur le média. |
| Texte | | Une séquence de caractères. |

Type — Un type est un ensemble nommé de caractéristiques. Ce nom peut être utilisé pour définir des ensembles de valeurs.

# 4. INTRODUCTION

## 4.1. OBJET DU DOCUMENT

Ce document est le résultat de l'étude menée par CS SI pour le compte du CELAR dans le cadre du thème 5 de l' « Etude sur les logiciels libres et la sécurité des systèmes informatiques ». Ce thème est dédié à la standardisation des formats de données.

L'étude proposée s'inscrit dans une démarche qui, à partir d'une description globale de la problématique liée à la standardisation des données, débouche sur la mise en œuvre pratique de standards pour les données de la DGA.

La première étape de la démarche vise donc à offrir une vue complète sur les enjeux de la standardisation des données et les moyens aujourd'hui disponibles. Nous proposons de développer les axes d'étude et de présentation suivants :

- *Sensibilisation à la problématique globale* de la standardisation des données (enjeux, objectifs) : l'approche est ici basée sur la connaissance pratique de CS SI sur le domaine et ses nombreux retours d'expérience dans les domaines scientifiques, aéronautiques et spatiaux. Parmi les points abordés, on peut noter : le cycle de vie des données, leur pérennisation, leur valorisation, les volumes croissants, …

- *Typologie de l'offre* : elle consiste en une recherche sur un ensemble de standards de données disponibles (normes, standards de fait, …). Pour chacun d'eux sont présentés leurs traits principaux, leurs domaines d'utilisation, l'outillage disponible. Les aspects liés aux logiciels seront en particulier analysés. Les tendances actuelles en terme d'utilisation et de mise en œuvre de ces formats sont aussi dégagées.

- *Présentation détaillée d'une norme* : L'implication de CS SI dans la définition et la mise en œuvre de la norme EAST permet d'offrir un éclairage particulier sur ce standard de données. Sont ainsi présentés donc de façon détaillée :

    - La norme EAST et son offre complémentaire : le standard DEDSL
    - Les outils permettant de mettre en œuvre cette norme sur le cycle de vie des données
    - Les applications cibles
    - Les principales références de mise en œuvre de la technologie (projets SPOT, HELIOS, ENVISAT, …)

## 4.2. PORTEE DE L'ETUDE

Au niveau de la portée de cette étude, il est utile de préciser que l'on ne traite que de formats de données et non pas de formats d'informations. Les formats de données s'attachent à définir des modes de représentation de données au sens « trains de bits », issues de mesures ou de traitements, et destinées à être analysées ou bien utilisées en entrée d'autres traitements. Ces données peuvent être issues de différents domaines applicatifs : technique, scientifique, télécommunication, gestion….

Les formats d'information sont ceux utilisés pour véhiculer des informations du type connaissance, vers des applications ou (la plupart du temps) des utilisateurs humains. Les standards de représentation des connaissances utilisés couramment sont des langages à balises du type «HTML », «SGML » et maintenant, «XML ». Ces types de formats ne sont donc pas analysés dans le cadre de l'étude.

Les formats de données abordés ne sont pas non plus des formats associés à des outils propriétaires. Ce sont des formats généralistes, non spécialisés, ne s'appliquant pas uniquement à des données du type «textes », « images », « multimédia ».

# 5. ENJEUX ET OBJECTIFS DE LA STANDARDISATION DES DONNEES

## 5.1. LES ENJEUX

L'industrie civile et militaire, les laboratoires de recherche ainsi que la plupart des acteurs étatiques ont à gérer des volumes de plus en plus importants de données scientifiques et techniques. On estime que dans la plupart des domaines, les volumes de données qui seront manipulés dans la prochaine décennie vont être de plusieurs dizaines à plusieurs centaines de fois plus importants que ceux manipulés aujourd'hui.

Les acteurs ont à résoudre des problèmes de définition, d'acquisition, de traitement, d'archivage et de distribution de ces informations. Ces problèmes surviennent dans des environnements de plus en plus complexes : les échanges d'informations se font entre des systèmes autonomes, distribués et le plus souvent hétérogènes. Ils nécessitent de plus un maximum de souplesse, si on prend en compte par exemple les modifications appliquées aux données entre le moment de leur génération et celui de leur application réelle.

En conséquence, dans un contexte de réduction des coûts et d'augmentation de la qualité et de la sécurité de ces données, le développement des standards de description des données, ainsi que des architectures et des systèmes permettant de les manipuler est un enjeu critique pour la plupart des organisations.

## 5.2. LES OBJECTIFS

La standardisation des données vise à fournir des moyens facilitant le développement d'environnements opérationnels et efficaces ou les utilisateurs impliqués dans le cycle de vie des données (cf. chapitre 5.3 ci-dessous) pourront :

- Comprendre ou décrire,
- Utiliser,
- Déterminer la disponibilité, la qualité,
- Demander la création, l'extraction, l'accès,
- Archiver,

des données (et métadonnées) qui auront ou pourront être crées par des entités (individus, expériences, organisations, …) et des environnements éventuellement étrangers à l'utilisateur.

Nous proposons dans cette étude de ne pas nous intéresser aux problèmes de l'archivage et de la restauration des données et de se concentrer sur les services associés à la compréhension et à la description des données, c'est à dire à la standardisation des **formats** de données. Ces services de haut niveau comprennent donc :

- La description des données par des descriptifs compréhensibles par l'être humain ou exploitables par des processus dédiés,
- L'association des données à leurs descriptifs,

- Le stockage, la recherche et l'administration des descriptifs,
- L'interprétation du contenu des données à partir des descriptions,

## 5.3. ANALYSE DE LA PROBLEMATIQUE

La problématique globale à laquelle se confrontent la plupart des organisations manipulant des données scientifiques et techniques peut être modélisée à l'aide du schéma suivant.

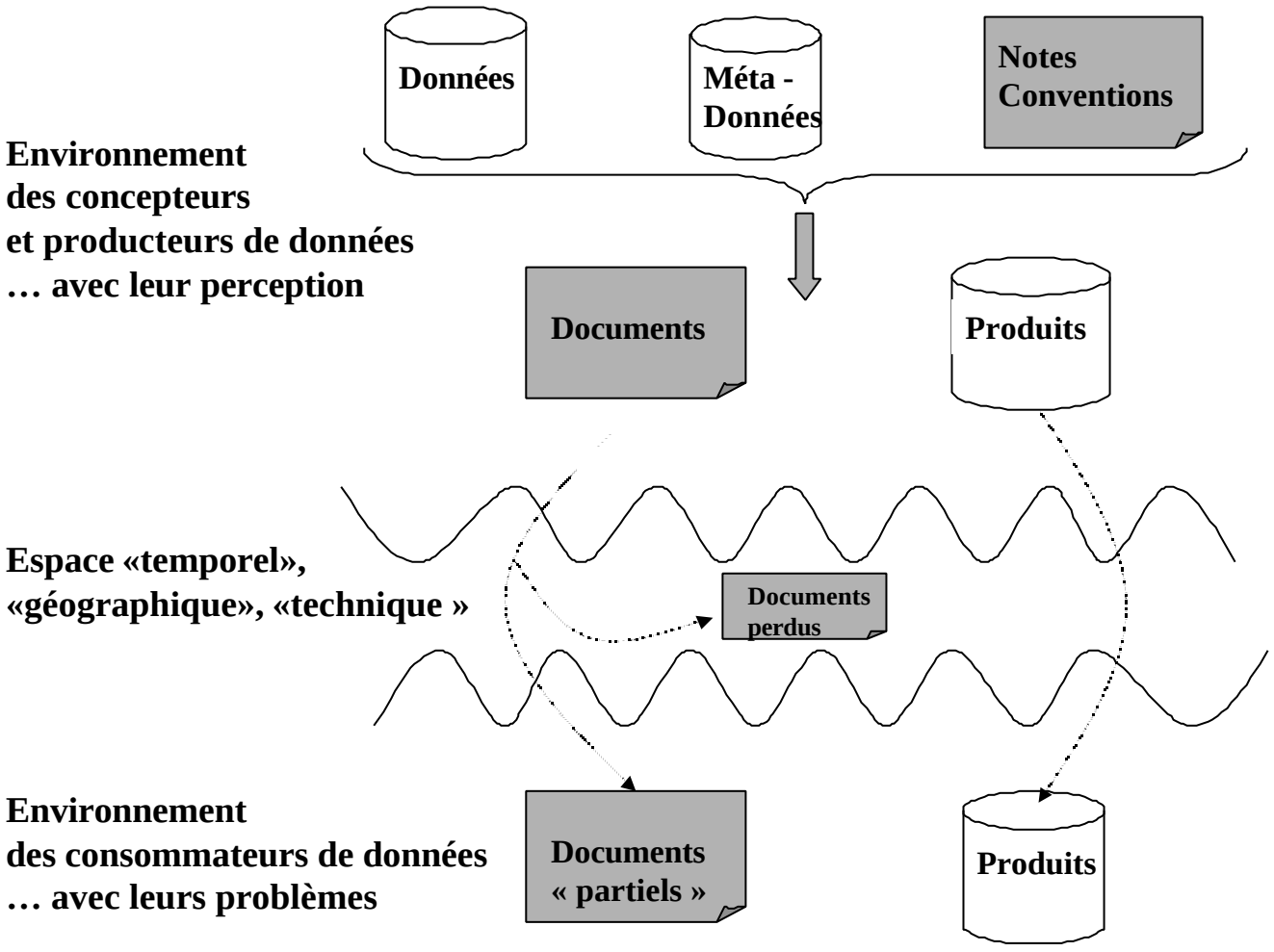


**fig. 1 : Problématique de la compréhension et de l'accès aux données**

Ce schéma recense les étapes clés du cycle de vie de la donnée, que nous proposons maintenant de parcourir afin d'analyser les modes de fonctionnement "standards".

**Conception et production des données**

Les **producteurs et concepteurs de données** (données elles-mêmes ou données accompagnant les données (i.e. métadonnées)) manipulent différents types de produits et ont une compréhension propre de leurs données. Cette compréhension est souvent documentée de façon partielle dans des notes, au mieux dans des documents. Ces documents peuvent inclurent des conventions (par exemple une définition de donnée par un type "maison") qui peuvent ne pas être documentées (que signifie ce type ?) et être même implicites. Ainsi, des informations comme les relations entre les différentes données peuvent ne jamais être renseignées.

Avant de rendre la donnée accessible aux utilisateurs "consommateurs", les producteurs "packagent" les données et en génèrent une description. Ce **document** vit la plupart du temps seul. Il peut faire référence à d'autres documents accessibles ou non aux futurs consommateurs. Le contenu du document est décidé par le producteur de données lui-même, selon la perception qu'il peut avoir du besoin de l'utilisateur. Il détermine le niveau de détail, les points de vue de la description, et l'organisation du document.

**Distribution des données**

Les données et la documentation associée peuvent être envoyées aux consommateurs des données ou bien dans des archives destinées à être exploitées ultérieurement. Le consommateur peut lui-même différer l'utilisation de la donnée. Le temps entre le moment ou la donnée est conçue et produite et le moment ou celle-ci est effectivement utilisée peut donc être propice à des pertes de documentation et donc d'informations.

La diffusion de données peut être effectuée par plusieurs moyens :

- Bandes magnétiques
- Disques optiques numériques
- Livraison par réseau (ftp, HTTP, ..)
- …

**Compréhension des données**

Lors de la réception de la donnée, il faut associer la donnée à la documentation et donc identifier la donnée pour trouver la bonne information. En effet, même les données s'appuyant sur des formats incluant leur propre description (voir chapitre 6 ci-dessous avec le format HDF par exemple), proposent rarement dans leur corps assez d'informations pour permettre aux scientifiques et autres utilisateurs d'analyser leurs contenus. Ces utilisateurs ont donc systématiquement besoin de la bonne documentation pour la donnée reçue. La correspondance entre le(s) document(s) et les données s'effectue alors de façon assez empirique, par des comparaisons entre les titres des documents et, par exemple, les étiquettes associées aux médias.

Une fois que la bonne documentation a été trouvée, le problème de sa compréhension se pose. L'hétérogénéité des documents proposés en fonction des différents formats impose des efforts sans cesse renouvelés pour comprendre le format et la signification des données. Il faut donc « apprendre » le dialecte propre au concepteur de la donnée et espérer que seulement un minimum d'informations implicites (celles faisant partie de l'environnement journalier du concepteur ou producteur) ont été oubliées dans le document.

**Exploitation des données**

Une fois la donnée comprise ou du moins une fois que le niveau des connaissances requis pour l'utilisation est atteint, le problème suivant est l'écriture de logiciels pour accéder à ces données. Même si la donnée est dans un format auto descriptif et que le consommateur possède les logiciels adéquats pour accéder à son contenu, des développements complémentaires sont la plupart du temps nécessaires pour extraire ou présenter des sous-ensembles de données plus adaptés aux besoins réels.

Dans cette thématique liée à l'exploitation des données, il faut prendre aussi en compte les problèmes liés à l'obsolescence des machines et à la diversité des intervenants d'un projet ou d'une expérience scientifique. Ainsi entre le moment où la donnée est conçue et celui où elle est accessible, distribuée ou exploitée, plusieurs années peuvent s'écouler. De plus, l'utilisateur cible pourra travailler sur des environnements radicalement différents de ceux où la donnée est produite. Ainsi les techniques de standardisation de données doivent fournir une description complète de la donnée aux niveaux logiques et physiques, afin de permettre l'exploitation de ces données indépendamment de la disponibilité du système original (matériel, système d'exploitation et logiciel) sur lequel elles ont été produites.

On pourra noter aussi l'importance de ce besoin par rapport à l'existence et à la constitution continue d'un véritable patrimoine technologique dans les différents domaines scientifiques que sont le spatial, le militaire ou bien encore le nucléaire. Aujourd'hui des archives extrêmement importantes sont en train de se constituer dans ces différents domaines. Il est fondamental de mettre en place une approche permettant de garantir non seulement la préservation physique des médias, mais aussi la préservation des informations et des moyens nécessaires à la restauration, la réhabilitation et l'exploitation des données composant ce patrimoine.

**En conclusion …**

On peut noter dans ces différentes phases des disparités au niveau des besoins des intervenants, de leurs perceptions, de leurs environnements. Le format de la donnée est l'information primordiale transitant dans le cadre de ces différentes étapes. Les disparités dans son contenu et dans sa forme sont à l'origine de la plupart des problèmes rencontrés. Si on s'intéresse au problème de la forme, on peut noter que celle ci peut prendre au cours des étapes du cycle de vie différentes formes pour traiter d'un problème unique :

- Le concepteur de la donnée la décrira à l'aide de textes en utilisant des éditeurs de texte standards (premier format),
- Il utilisera des notations inspirées de langages de programmation pour décrire leur organisation sur les médias (deuxième format),
- Le producteur de la donnée utilisera un langage informatique pouvant être différent pour générer ses produits (troisième format). Ces formats ne seront pas la plupart du temps livrés avec le produit,

- L'utilisateur devra à son tour décrire les données à partir des documents livrés, dans un langage pouvant être aussi différent, et avec sa propre interprétation des documents descriptifs (quatrième format).

Un référentiel unique, commun à l'ensemble des intervenants sur le cycle de vie de la donnée, support à la description, à la production et à l'utilisation des données, constituerait donc une approche cohérente.

## 5.4. LES APPORTS DE LA STANDARDISATION DES DONNEES

Les apports de la standardisation des données se situent au niveau des trois problèmes fondamentaux que sont :

1. La correspondance entre les données et leur documentation
2. La compréhension des documents associés aux données
3. Le besoin de développer des logiciels d'accès pour chaque nouvelle donnée

Au **problème de la correspondance entre les données et leur documentation**, la standardisation peut répondre en créant des mécanismes standards pour lier de façon non ambiguë les structures de données avec leurs descriptifs. Cette approche peut être menée de deux façons différentes, soit en intégrant la description à la structure de donnée, soit en la séparant.

La première approche possède le gros avantage d'avoir cette description de façon immédiate. Elle résout donc immédiatement le problème de la correspondance. Quelle que soit la donnée on a la certitude d'accéder à son contenu. Elle possède cependant quelques sérieux désavantages :

- Elle impose une structure aux données et donc contraint les possibilités de contenu et de structure.
- La description est limitée aux possibilités du format et nécessite souvent des informations (donc des documentations) complémentaires.
- Elle ne peut pas traiter de données existantes, sauf si pour ces données un effort de transfert du format initial vers le format intégré a été effectué
- Elle fragilise la donnée, liant sa pérennité à celle des produits supports.

La deuxième approche nécessite des méthodes de packaging et d'assignation :

- Le packaging va constituer en la définition de structures de haut niveau associant les produits avec leurs descriptifs.
- L'assignation va consister par exemple à affecter un identifiant unique à chaque description. Cet identifiant est associé à la donnée. Il faut donc mettre en place un standard de structure pour gérer ces identifiants, enregistrer, maintenir et distribuer les descriptions de données. Une autorité de contrôle doit donc être crée pour supporter cette infrastructure (autorité internationale, nationale, propre à l'entreprise, à la profession, …).

On voit donc ici que la standardisation des données impacte potentiellement sur les structures des organisations. Cette approche, souvent nécessaire est lourde à analyser et déborde le champ d'application de la présente étude. Il nous semblait cependant nécessaire de l'évoquer afin de traiter, à ce niveau de l'étude, le problème dans sa globalité.

Au **problème de la compréhension des documents associés aux produits**, la standardisation peut répondre en créant des standards pour l'écriture des descriptions de données. Les descriptions des données peuvent être de types plus ou moins complexes. On peut en effet aller de la description de structures de bits pour des types de données simples (entiers, réels,..) à la description globale de mesure d'instruments.

Ces informations doivent être compréhensibles par le consommateur, qu'il soit le lecteur «humain » ou bien le processus chargé de l'exploitation de ces données. Les langages disponibles ne possèdent pas les caractéristiques permettant la fourniture simultanée de la description de la donnée en langage proche du langage naturel ainsi que la possibilité d'être interprétés automatiquement par des processus. Ainsi un ou plusieurs langages doivent être combinés pour obtenir la capacité de compréhension et d'automatisation.

On trouvera donc au niveau de la standardisation des données des approches permettant :

- La description des formats à travers des langages interprétables par les machines
- La production de dictionnaires de données pour la compréhension des produits
- La mise en place de liens associant les noms des champs des produits non seulement à leurs valeurs mais aussi à leurs descriptions.

Au **problème consistant à développer des logiciels d'accès pour chaque nouveau produit,** la standardisation peut répondre en offrant des formats indépendants des différentes disciplines ou ils s'appliquent et supportant l'automatisation des accès aux données (analyse, ingestion ,…). Parmi ces fonctions automatisées, on peut trouver :

- l'identification d'objets de données sur le média en cours d'exploitation,
- l'aptitude à décomposer ces objets de données en sous objets et ainsi de suite jusqu'aux données élémentaires,
- la possibilité de présenter les significations de ces objets aux consommateurs,
- la capacité à comprendre les relations entre les objets de données.

# 6. STANDARDS DE DONNEES : TYPOLOGIE DE L'OFFRE

Les standards relatifs aux données sont nombreux : standards relatifs à l'empaquetage des données, à la description des données ou bien encore au codage des temps et des dates.

Tous ces standards ne se situent pas au même niveau et ne sont pas tous compatibles. Certains imposent une structure spécifique aux données (FITS utilisé en Astronomie et en Physique solaire, CDF utilisé dans le programme international ISTP, HDF utilisé dans les Sciences de l'Environnement), d'autres ne font aucune hypothèse sur cette structure ( SFDU, EAST). Dans tous les cas, leur mise en œuvre constitue toujours un progrès par rapport à l'absence de tout standard.

Nous proposons dans ce chapitre une typologie de l'offre dans le monde des formats de données scientifiques et techniques. En balayant les caractéristiques fines de cette offre par rapport à la problématique étudiée au chapitre précédent, le chapitre tentera de souligner les tendances fortes. Ce chapitre ne s'intéresse qu'à des formats de données et non pas à des standards de description de données. Les standards de description des données seront étudiés de façon très précise dans le chapitre suivant. On pourra ainsi, en conclusion, effectuer un parallèle entre ces deux offres ou approches.

Pour les formats représentatifs analysés la présentation comprendra en dehors des traits principaux, les domaines d'utilisation ainsi que l'outillage disponible (aspect logiciels libres).

Les formats seront resitués par rapport aux apports essentiels de la formalisation des données (lien des données avec la documentation, compréhension des produits, logiciels d'accès).

Un tableau récapitulatif permet d'offrir un aperçu global de l'offre.

## 6.1. LE FORMAT CDF

### 6.1.1. DESCRIPTION DU FORMAT

#### *6.1.1.1. GENERALITES*

Comme la plupart des formats que nous présentons dans ce document, CDF (***Common Data Format***) se présente comme étant un format auto-descriptif de données basé sur le concept de l'abstraction des données. Les données sont conceptuellement représentées sous forme de tableaux multidimensionnels.

Le terme « auto-descriptif » signifie qu'en plus de contenir les données réelles, c'est à dire les valeurs scientifiques, le format contient également des informations qui décrivent ces données. Ces informations renseignent sur la signification des données mais aussi sur leur organisation, ce sont les métadonnées. Cette auto-description permet au format de couvrir des champs d'applications scientifiques divers.

Le terme « abstraction des données » signifie que les données sont connues :

- d'une part, par la représentation conceptuelle qui en a été faite,
- et d'autre part, par les moyens fournis pour manipuler cette représentation. Ces moyens sont fournis par un ensemble d'outils logiciels, librairies sous forme d'interface logicielle.

Ceci veut dire que l'utilisateur n'a pas besoin de connaître comment les données et les métadonnées sont physiquement stockées (la notion de bits et d'octets est occultée).

Le terme «CDF » est utilisé dans la suite de ce chapitre pour faire référence aux fichiers physiques générés comme aux logiciels CDF associes.

#### *6.1.1.2. COMPOSITION DU FORMAT CDF*

Un format CDF est composé d'un ensemble de variables (les data) et d'un ensemble d'attributs (les métadata).

L'ensemble de variables s'appelle un «CDF record » et constitue les **objets de données**. Un CDF peut contenir, et c'est généralement le cas, plusieurs CDF records.

Une variable est un scalaire, un vecteur ou plus généralement un tableau à n dimensions. Le nombre de dimensions et la taille de chaque dimension d'une variable dépend des données et sont définis par le concepteur de ces données (la limite étant n=10). Une variable peut être de deux types : les rVariables dont les dimensions sont fixes, les zVariables dont les dimensions varient.

Des tables de variance sont ensuite établies, qui spécifient comment évoluent les valeurs des variables dans le CDF (cycle de données, répétitions d'observations scientifiques a intervalle de temps régulier) afin d'optimiser leur nombre.

Un grand nombre de données de disciplines scientifiques variées peuvent ainsi être agencées sous forme de tableaux multidimensionnels. Cela nécessite d'avoir de l'expérience dans la matière et de bien connaître les dépendances et les corrélations entre les données. La présence d'un expert au niveau de la compréhension des données est nécessaire pour formaliser l'agencement adéquat, et ainsi concevoir les données CDF.

L'ensemble des attributs (métadonnées) décrit globalement un CDF ou décrit spécifiquement une variable CDF. L'ensemble des attributs d'un CDF constitue le **dictionnaire des données**. On distingue les gAttributs et les vAttributs.

Les gAttributs sont utilisés pour décrire le CDF dans sa globalité ou pour décrire des propriétés communes a toutes les variables, par exemple pour donner un titre au CDF ou préciser son historique (date de création, date de chaque modification) ou pour tout simplement documenter textuellement le CDF.

Les vAttributs sont utilisés pour décrire les propriétés d'une variable en particulier, par exemple pour spécifier le minimum et le maximum des valeurs d'une donnée ou pour lui donner un nom.

Il est possible de définir autant de gattributs ou de vAttributs qu'il est nécessaire. Pour chaque nouvelle information incluse dans un attribut, une gEntrie ou une vEntrie est définie. Le dictionnaire des données n'est donc pas fixé par le format et une sémantique ciblée a un domaine scientifique particulier peut être définie.

#### *6.1.1.3. ORGANISATION DU FORMAT CDF*

Un format CDF peut être organise de deux façons différentes :

- soit dans un fichier unique cdf contenant les données et les métadonnées. L'avantage de l'organisation en un seul fichier est de n'avoir à gérer qu'un fichier unique(pas de soucis de pertes d'un fichier surtout dans les transferts réseaux). Par contre, l'accès aux données peut devenir très vite complexe et les temps de traitement beaucoup plus longs.

- soit dans plusieurs fichiers. Dans ce dernier cas, il existe un fichier cdf contenant les métadonnées et les informations de contrôle sur les données. Associé à ce fichier cdf, pour chaque variable définie dans le format, il existe un fichier de données. Cette organisation multi-fichiers, délimite clairement les données et les métadonnées, ce qui facilite les mises a jour, les ajouts, et rend les accès aux données plus rapides. Par contre, le transfert des données comporte plus de risques.

### 6.1.2. LOGICIELS CDF ASSOCIES

CDF possède deux niveaux d'accès aux données. L'un des niveaux d'accès s'effectue en s'appuyant directement sur l'interface programmatique de la librairie CDF. L'autre niveau s'effectue au travers d'outils CDF eux-mêmes écrits en s'appuyant sur la couche de l'interface programmatique. Chacun des niveaux est dédié à une classe d'utilisateurs, le niveau interface programmatique étant dédié aux développeurs d'applications spécifiques, le niveau outils étant dédié aux concepteurs et aux créateurs de données.

#### 6.1.2.1. LA LIBRAIRIE CDF

Elle se compose de deux interfaces : la *Standard Interface* comprenant trois groupes de fonctions manipulant le CDF : fonctions sur les attributs, fonctions sur les variables, et fonctions générales sur le HDF. Cette interface s'appuie sur une interface de plus bas niveau, « l'*Internal Interface* ».

Elle permet aux développeurs de systèmes basés sur le format CDF d'écrire facilement des applications qui traitent ces données CDF, comme extraire des sous-espaces multidimensionnels de données, accéder à des structures complètes, échantillonner les données ou bien isoler une donnée particulière. Toutes ces manipulations sont possibles sans que l'utilisateur ait besoin de connaître la représentation physique des données. Il est dégagé des problèmes de programmation des couches de bas niveaux concernant les E/S. Il doit à la création des données spécifier l'encodage désiré. Les données CDF peuvent être ensuite lues, décodées et converties dans un encodage correspondant à une autre machine.

Un autre aspect de la librairie est la possibilité de compresser les données lors de l'écriture sur disque et de les décompresser lors de la lecture, ceci aussi de façon transparente pour l'utilisateur. Il est possible de compresser un CDF entier ou de ne compresser qu'une variable (un CDF multi-fichier ne peut pas par contre être compressé). Le logiciel de compression peut être choisi par l'utilisateur.

Cette librairie s'accompagne d'une interface programmatique applicative C, Fortran et dernièrement Java.

#### 6.1.2.2. LES OUTILS CDF

Les outils CDF offrent surtout des services pour créer de nouveaux CDFs et pour consulter les CDFs existants. Ils permettent de structurer un CDF et de décrire les métadonnées sans utiliser l'interface programmatique. Les outils disponibles sont cites ci-après :

- CDFedit permettant l'affichage, la création et la modification des variables et des attributs d'un CDF,
- CDFexport permettant le transfert d'une partie d'un CDF a l'écran, dans un fichier texte ou dans un autre fichier CDF.
- CDFconvert permettant de changer l'encodage, la compression d'un CDF, l'organisation d'un CDF...,
- CDFcompare pour comparer deux CDF entre eux,
- CDFstats pour produire des résultats statistiques sur les variables CDF,
- SkeletonTable qui crée un squelette CDF à partir d'un fichier texte appelé skeleton table,
- SkeletonCDF qui crée un fichier texte appelé skeleton table à partir d'un CDF et peut être ensuite utilise après modification pour créer un nouveau CDF,
- CDFInquire affiche des informations sur la version CDF.

#### 6.1.2.3. ACCES AUX LOGICIELS

Les données CDF ainsi que les logiciels CDF associés sont portables sur les plates-formes les plus usuelles. Le tableau ci-dessous en donne la liste :

| Plates-Formes | OS |
|---|---|
| DEC Alpha | OSF/1, OpenVMS |
| DECstation | Ultrix, VMS |
| HP 9000 series | HP-UX |
| PC | MS-DOS/Windows 3.x/95/98/NT, Linux, QNX |
| IBM RS6000 series | AIX |
| Macintosh | MacOS 7.0 |
| NeXT | Mach |
| SGI Iris, Power series, Indigo | IRIX |
| Sun | SunOS, SOLARIS |
| VAX | VMS |

Les logiciels CDF sont disponibles via FTP anonyme et DECnet pour les systèmes OpenVMS, via FTP anonyme seulement pour les systèmes UNIX/POSIX shell ainsi que les systèmes MS-DOS et Macintosh.

Des informations générales sur CDF et ses logiciels associés sont disponibles sous l'URL http://nssdc.gsfc.nasa.gov/cdf/.

### 6.1.3. LES DOMAINES D'APPLICATION

CDF a été créé au NSSDS (National Space Science Data Center). Cette organisation a pour vocation de gérer les accès aux données provenant de diverses missions spatiales de la NASA. Ce format a été élaboré avec la participation de l'organisme NOST (NASA/Science Office of Standards and Technology). Cet organisme a été créé pour encourager l'évolution et l'adoption de standards de données pour faciliter les échanges de connaissance entre communautés scientifiques. Il est présent dans les activités ISO et CCSDS.

Les utilisateurs de CDF se situent dans des domaines assez variés : il est utilisé par plusieurs organismes gouvernementaux, universités, sociétés privées et commerciales, et certains centres de recherche. Cependant, une des principales utilisations se situe dans la physique de l'espace.

Ainsi, CDF a été choisi pour les programmes de l'ISTP (International Solar-Terrestrial Physics) et par l'IACG (Interagency Consultative Group ) dont l'objectif est l'étude des plasmas planétaires et interplanétaires. Dans ce cadre la, un guide d'utilisation standard ISTP/IACG CDF (*guidelines ISTP*) a été établi et de nombreux outils ont été développés dans le but de créer, puis d'analyser un ensemble de données ISTP/IACG CDF, et enfin de visualiser les résultats de ces analyses, de façon à les rendre directement interprétables par les scientifiques ( courbes, tracés graphiques, spectrogrammes, images 3D...), sans exiger, ou peu, de connaissances de leur part quant à la formalisation même de ces données.

Citons, en exemple, l'outil KPVT (ISTP Key Parameter Visualization Tool) qui, comme la plupart de ces outils, est écrit en IDL, présente une interface utilisateur graphique, et est portable sur les plate-forme les plus usuelles (Stations Unix, VAX/OpenVMS, PC/Windows et Macintosh). L'ensemble de ses codes sources, les instructions d'installation peuvent être obtenus via FTP anonyme ou à partir du site web ISTP.

En parallèle, le CDAWeb (Coordinated Data Analysis Web) possède une base de données ISTP/IACG CDF multi-instruments et multi-missions. Sa fonction consiste en la diffusion de données et en la diffusion des logiciels d'analyses associes. Certains sont publics, d'autres requièrent un mot de passe pour y accéder (http://nssdc.gsfc.nasa.gov/spdf/sp_use_of_cdf.html)

Nous citons ici quelques autres applications développées avec CDF :

- CWIT CDF Windows Imagind Tool (disponible sur MS-Windows, Macintosh et UNIX).
- VirtualCOHO , COHOWeb (visualisation 3-D d'images spatiales provenant de diverses missions (Helios, Ulysse, Pioneer)).

## 6.2. LE FORMAT HDF

### 6.2.1. DESCRIPTION DU FORMAT

#### *6.2.1.1. GENERALITES*

En s'appuyant sur les termes « auto descriptif » et « abstraction de données » développés au paragraphe précédent, le format HDF (***Hierarchical Data Format***) se caractérise essentiellement, tout comme CDF, comme étant un format auto descriptif basé sur le concept de l'abstraction des données. Ce format permet de stocker et de gérer des données de type texte, numérique ou graphique.

#### *6.2.1.2. COMPOSITION DU FORMAT HDF*

La structure générale HDF consiste en un index d'étiquettes, reliées entre elles, chacune d'elle décrivant les données et les localisant.

Un fichier HDF comprend une entête (*file header*), suivie d'au moins un bloc descripteur de données (*data descriptor block*), suivie de zéro ou plusieurs éléments de données (*data elements*).

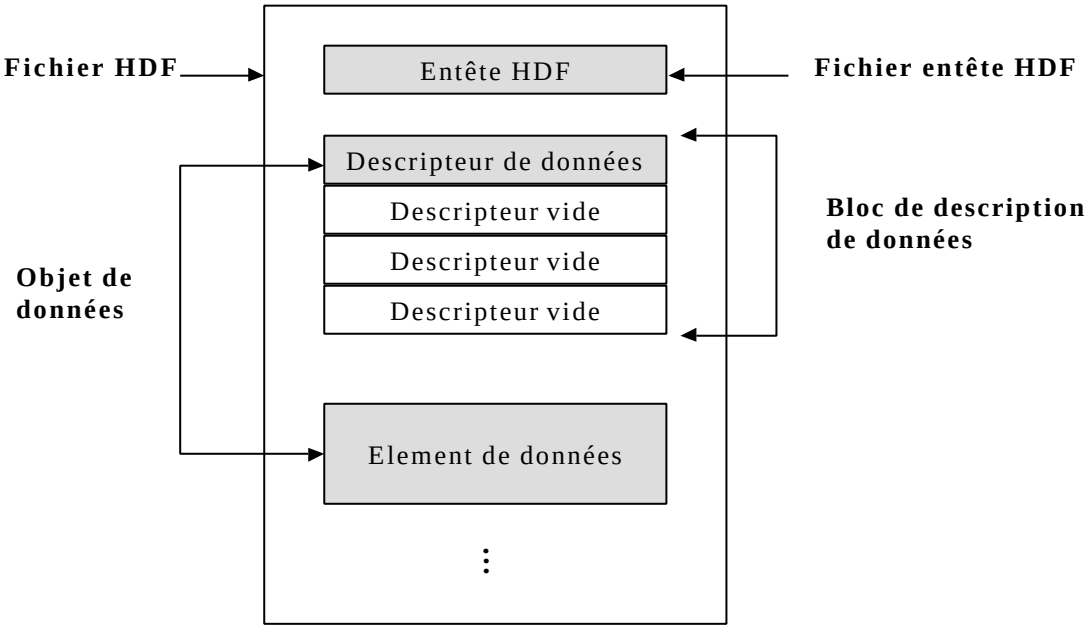


**fig 2 : Structure d'un fichier HDF**

- Le f*ile* h*eader* (entête HDF) indique que le fichier est au format HDF (4 octets)

- Un *data descriptor block co*ntient plusieurs descripteurs de données (data descriptor). Un descripteur de données est associe a un élément de données (*data element*) et forment ensemble **un objet de données** (*object data*) qui est la structure de base du format HDF. Les *data descriptors* sont relies entre eux par pointeurs. De plus, le *data descriptor block* possède une entête (*data descriptor header*) composée du nombre de *data descriptors* (par défaut 16) qu'il contient (*block size*) , et de l'emplacement du prochain *data descriptor block (next block).*

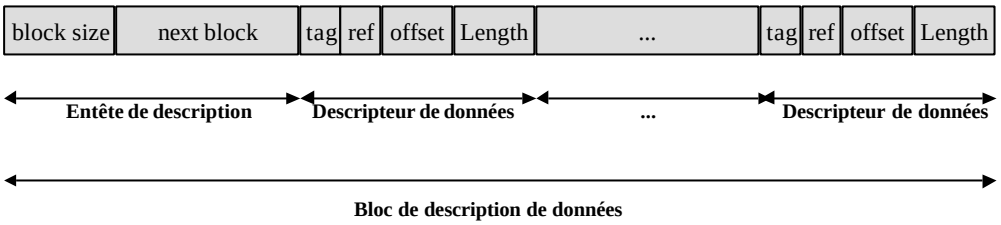


**fig. 3 : Bloc de description de données HDF**

- Un *data descriptor* contient des informations sur le type , la localisation et la taille de l'élément de données correspondant. Il est de taille fixe (12 octets) et est constitue de la manière suivante:

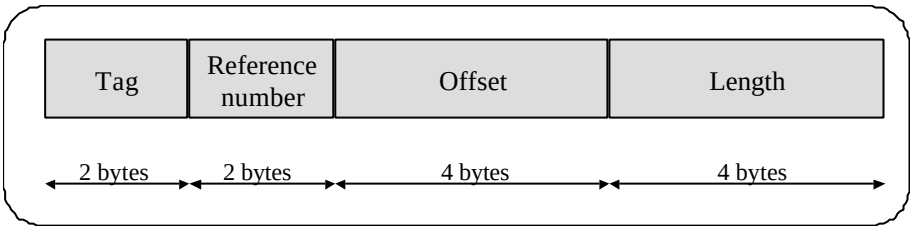


**fig. 4 : Descripteur HDF**

  - « TAG » indique le type de base de la données (il en existe 6) contenues dans le *data element* et est associé à un mnémonique pour améliorer la lisibilité des programmes HDF.
  - La paire TAG et « reference number » identifie de manière unique le *data object*.
  - Le champ *data offset* fournit l'octet de début du data élément par rapport au début du fichier.
  - Le champ *length* contient la taille en octets du *data element*.

- Un élément de données (*data element*) contient les données « pures » de l'*object data.*

Il existe six types de base pour les données :

- 8-bit raster images et 24-bit raster images: ils correspondent aux images raster sur 8 et 24 bits comme leur nom l'indique.
- palette : il s'agit du type de données servant à représenter une palette de couleurs associée a une image.
- scientific data model (SD): permet de stocker des tableaux multi dimensionnels d'entiers ou de réels.
- VData : il s'agit de tableaux dont chaque élément est un record constitué d'entiers, de réels ou de caractères.
- Annotation Model : ce sont des descriptions textuelles associées aux différents éléments du fichier HDF.

Les *data objects* ayant un lien entre eux sont regroupes en *data sets* appelés *Vgroup* (un *Vgroup* peut contenir par exemple une image raster et sa palette associe). Sachant qu'un Vgroup peut être constitue d'un ensemble de Vgroups, la notion de données hiérarchisées se justifie ici.

#### *6.2.1.3. ORGANISATION DU FORMAT HDF*

Un format HDF peut être stocké physiquement :

- soit sous un fichier unique,
- soit, dans les versions HDF les plus actuelles, sous plusieurs fichiers partageables.

### 6.2.2. LOGICIELS HDF ASSOCIES

#### *6.2.2.1. LES LIBRAIRIES ET LES OUTILS HDF*

HDF peut être vu sous plusieurs niveaux :

- Au niveau le plus bas, HDF est représenté par les fichiers physiques au format HDF,
- A un niveau supérieur, HDF est un ensemble d'utilitaires ou *APIs* pour manipuler, visualiser et analyser les données contenues dans les fichiers. Les services que ces APIs offrent, sont accessibles depuis une application C ou Fortran. On distingue deux catégories d'APIs suivant l'organisation des fichiers ; l'interface multi-fichiers manipulant des HDF multi fichiers et l'interface mono fichier manipulant des HDFs mono fichier. Cette dernière n'étant pas compatible avec la version plus récente multi fichiers, les deux interfaces continuent d'exister en parallèle. Chaque API offre un ensemble de services pour chaque type de base (Image Raster, Palette, SD...).
- Entre ces deux niveaux, la librairie HDF fait le lien entre les fichiers physiques et les APIs. Elle règle les problèmes tels que les E/S sur fichiers, la gestion des erreurs, la gestion de la mémoire. Ce niveau d'accès est réservé aux développeurs d'applications HDF particulières.
- Au niveau le plus haut, *applications générales*, HDF est un ensemble d'utilitaires en ligne (commandes en ligne lancées comme des commandes shells sous Unix) permettant de réaliser des manipulations sur les fichiers HDF sans avoir à les programmer. Elle comprend aussi certaines applications spécifiques NCSA, tel que JHV (Java-based HDF Viewer) qui est un outil de développement d'applications basées sur HDF.

Cette organisation en niveau peut être représentée grâce à la figure ci dessous.

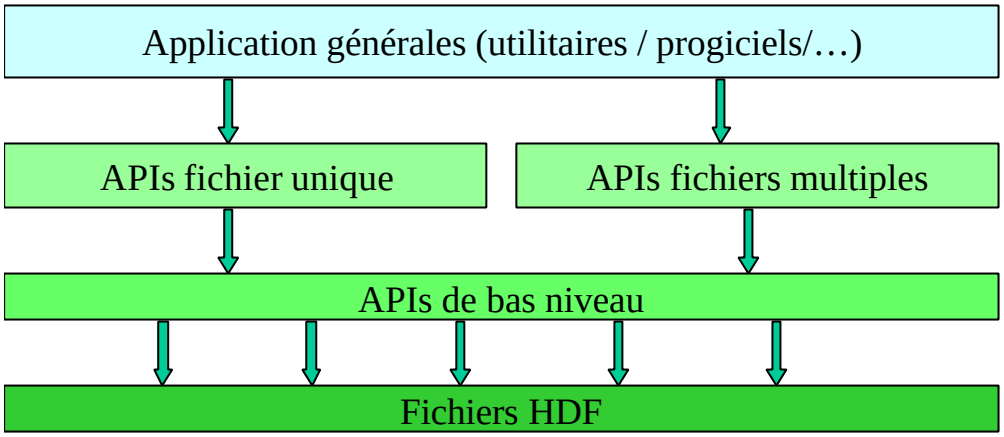


**fig. 5 : Les trois niveaux d'accès aux fichiers HDF**

Nous listons, dans la suite, les APIs existantes et quelques commandes en ligne. Chacune d'elles regroupe un ensemble de fonctions manipulant un type de base CDF particulier.

Les APIs HDF multi fichiers se composent de :

- SD API pour stocker, gérer et extraire des tableaux multi dimensionnels de caractères ou de valeurs numériques,
- VS API pour stocker, gérer et extraire des données de types différents agencées sous forme de structures dans un tableau,
- V API pour créer des groupes de structures HDF,
- GR API pour stocker, gérer et extraire des images raster ainsi que leur palette de couleurs associées,

Les API s HDF simple fichier se composent de :

- DFR8 API pour stocker, gérer et extraire des images raster sur 8 bits ainsi que leur palette de couleurs associées contenues dans un fichier unique,
- DFR24API pour stocker, gérer et extraire des images raster sur 24bits ainsi que leur palette de couleurs associées contenues dans un fichier unique,
- DFP API pour stocker, gérer et extraire des palettes de couleur 8 bits dans un fichier unique,
- DFAN API pour stocker, gérer et extraire des textes de caractères décrivant un fichier ou une structure de données particulière.
- DFSD API pour stocker, gérer et extraire des tableaux multi dimensionnels d'entiers ou de réels suivant leurs dimensions et leurs attributs.

Ci-dessous sont listées quelques commandes en ligne :

- *hdp* (HDF dumper) affiche des informations générales sur le HDF,
- *vshow* affiche les informations concernant un data set,
- *fp2hdf, r8tohdf, r24hdf8, paltohdf* permet la conversion de type basique «brut »(réels, images raster 8-bit ...) en type de base HDF,
- *hdftor8, hdftopal* permet la conversion de type de base HDF en type basique « brut »,
- *ristosds, hdf24hdf8* permet la conversion d'un type de base HDF vers un autre type de base HDF,
- *hdfcomp* permet la compression d'images raster 8-bit,
- *hdfpack* permet la compression d'un fichier HDF.

#### *6.2.2.2. ACCES AUX LOGICIELS*

Les données HDF ainsi que les logiciels HDF associés sont portables sur les plates-formes les plus usuelles. Le tableau ci-dessous en donne une liste exhaustive :

| **Plates-Formes** | **OS** |
|---|---|
| Sun Sun4 | SunOS, Solaris |
| SGI Indy, PowerChallenge, Origin | Irix |
| H/P HP9000 | HPUX |
| SGI/Cray | UNICOS |
| DEC Alpha | Digital Unix, OpenVMS |
| DEC VAX | OpenVMS |
| PC | Solaris86, Linux, FreeBSD |
| PC | Windows NT/95 |
| Apple Power Macintosh | MacOS |

Les codes sources et la documentation des librairies HDF et des APIs, ainsi que les binaires dédiés à chaque plate-forme supportant HDF sont libres d'accès mais soumis aux restrictions de copyright. Ces sources et documents sont accessibles via le serveur anonyme FTP du NCSA. Des applications NCSA, ainsi que des applications écrites par des membres de la communauté des utilisateurs HDF sont aussi disponibles via ce serveur.

### 6.2.3. LES DOMAINE D'APPLICATION

HDF a été créé au NCSA (National Center for Supercomputing Application) a l'université de l'Illinois, à partir de besoins identiques recenses par des communautés scientifiques diverses. Les champs d'applications sont assez varies (expériences spatiales, observations océaniques, climatiques, environnement de la terre, ...).

On peut citer parmi les organisations qui l'utilisent :

- ACE (Advanced Composition Explorer) spacecraft dans une mission d'observations des particules d'énergie du système solaire,
- Aerodyne's Center for Optical Signature Recognisation ont utilise HDF dans un premier temps comme moyen de transfert de données d'une application a une autre,
- ARM -Atmospheric Radiation Measurement Program,
- ASCI- Accelerated Strategic Computing Initiative,
- The Atmospheric and Oceanic Group au NCSA,
- Global Aerospace Corporation,
- GE Aircraft Engines,
- Information Technology and Systems Center (ITSC) en relation avec la NASA a intégré des données spatiales (images satellites) sous HDF, et a participe à l'élaboration du format HDF-EOS,
- The Odin Satellite Project pour des données d'une expérience relative à l'ozone à partir du satellite Odin,
- PFEL (Pacific Fisheries Environmental Laboratory)
- NASA's Distributed Active Archive Center (DAAC)

## 6.3. LE FORMAT FITS

### 6.3.1. DESCRIPTION DU FORMAT

#### *6.3.1.1. GENERALITES*

FITS (***Flexible Image Transport System***) est un format de données conçu pour faciliter les échanges de données d'astronomies entre plusieurs installations dont les formats internes diffèrent et pour donner un format standard pour archiver ces données.

#### *6.3.1.2. COMPOSITION DU FORMAT FITS*

Un fichier de données FITS est compose d'une séquence de HDUs ( *Header + Data Unit*). L'entête (*header*) est composée d'instructions ASCII « *keyword=value/commentaires* », qui décrivent l'organisation des données dans le HDU ainsi que leur format. Les données, *Data Units*, suivent ensuite dans le HDU, structurées comme indique dans l'entête.

Le premier HDU s'appelle le « PRIMARY ARRAY » et est un tableau de n dimensions. Les HDUs suivants sont appelés des « extensions », et sont de trois types différents :

- Image Extension qui est un tableau de n dimensions de pixels comme le PRIMARY ARRAY,
- ASCII Table Extension qui est un tableau lignes colonnes de caractères ASCII,
- Binary Table Extension qui est un tableau lignes colonnes de données en représentation binaire.

FITS supporte cinq types pour représenter ses données élémentaires : les entiers 8-bit non signés, les entiers 16-bit signes complément à 2, les entiers 32-bit signes complément à 2, les réels 32-bit IEEE, les réels 64-bit IEEE.

Chaque header ou Data Unit doit comporter exactement 2880 octets.

Chaque instruction (*keyword=value/commentaires)* doit comporter exactement 80 caractères ( le format de chaque instruction est fixe dont 8 caractères pour le mot clé). Celles qui précisent la taille et le format des données sont obligatoires. D'autres sont optionnelles, comme « COMMENT » ou « HISTORY » et permettent de documenter les données. La dernière instruction d'un HDU comporte le mot clé « END » et ne comporte ni de valeur ni de commentaire.

### 6.3.2. LOGICIELS FITS ASSOCIES

#### *6.3.2.1. LES LIBRAIRIES ET LES OUTILS FITS*

FITS ne possède pas encore de logiciels standards génériques. Les utilisateurs doivent développer leurs propres logiciels pour lire et visualiser leurs données. Il existe tout de même un nombre de paquetages développés pour des applications particulières. Pour en citer quelques-uns uns, citons ceux développés par l'HEASARC (High Energy Astrophysics Science Archive Research Center) de la NASA regroupés sous HEASOFT qui est un ensemble de produits logiciels intégrés utilisant les mêmes bibliothèques de base. Il comprend :

- FITSIO qui est un paquetage de lecture, écriture et modifications de fichiers FITS,

- FTOOLS qui est un paquetage de divers utilitaires. Il contient des utilitaires généraux pour manipuler les fichiers FITS ainsi que des programmes d'analyses plus sophistiques dédiés à des missions d'astrophysique bien spécifiques. FTOOLS est portable sur ALPHA/OSF, DEC/Ultrix, Linux, SUN/SunOS, SUN/Solaris, HP/HP-UXm, SGI/IRIX,VMS: ALPHA/VMS, VAX/VMS.

- XANADU qui est un programme multi missions d'analyse de données spectrales.

- XSTAR qui est un outil de calculs des conditions physiques et d'émissions spectrales de gaz photo ionisés

#### *6.3.2.2. ACCES AUX LOGICIELS*

HEASOFT est distribue sous forme de binaires ou sous forme de sources (avec documentation associée). FTOOLS, XANADU sont accessibles à partir de http://heasarc.gsfc.nasa.gov/lheasoft ou par ftp://legacy.gsfc.nasa.gov/software/lheasoft. Ils peuvent être téléchargés et installés ensemble ou séparément.

FITSIO est distribue via ftp anonyme à partir de ftp://heasarc.gsfc.nasa.gov/software/fitsio.

### 6.3.3. LES DOMAINE D'APPLICATION

FITS est un format de données standard approuve par NOST (NASA/Science Office of Standards and Technology). Il existe un service support (le FITS Support Office Software) responsable de la documentation du standard, qui participe a l'évolution de ce dernier et conseille sur la façon de concevoir les données d'une mission d'astrophysique particulière de la NASA.

Son domaine d'application est essentiellement l'astronomie et l'astrophysique. Dans ce cadre là, un groupe de travail, le IAUFWG (International Astronomical Union FITS Working Group) s'est formé auquel sont soumises toutes propositions d'évolution du format. Celui-ci décide de refuser ou d'accepter l'évolution en prenant comme règle première, dans le cas de l'acceptation, d'assurer toujours une compatibilité avec les versions antécédentes.

## 6.4. APERÇU GLOBAL DE L'OFFRE

Le tableau suivant récapitule les caractéristiques essentielles des formats de données étudiés :

| Format | Type | Liens Données / Documentation | Logiciels d'accès | Liens |
|---|---|---|---|---|
| **CDF** | Un seul type de données: tableaux n dimensions.<br><br>Implémentation sur un ou plusieurs fichiers. | Auto-documentation axée sur la sémantique des données par l'utilisation d'attributs.<br><br>Aucune documentation sur l'organisation des données.<br><br>Métadonnees : peuvent être dans un fichier séparé des données | Deux niveaux d'accès:<br><br>-outils stand-alone<br><br>-Interfaces de bibliothèques C, Fortran (et Java). | http://nssdc.gsfc.nasa.gov |
| **HDF** | Plusieurs types de base(6) regroupés et hiérarchisés (Vgroup).<br><br>Contraintes sur les formats des descripteurs.<br><br>Implémentation sur un ou plusieurs fichiers. | Auto-documentation axée sur l'organisation des données par les descripteurs.<br><br>Peu de sémantique. | Trois niveaux d'accès<br><br>-applications NCSA, commandes en ligne<br><br>-Ensemble d'APIs C et Fortran génériques.<br><br>-bibliothèques | http://hdf.ncsa.uiuc.edu |
| **FITS** | 3 types de base.<br><br>Fortes contraintes sur le format du header (instructions).<br><br>Contraintes sur la taille des données. | Auto-documentation axée sur l'organisation des données par les headers.<br><br>Peu de sémantique (mots clés sont limités).<br><br>Sémantique figée (mots clés sont prédéfinis). | Pas de logiciels standards d'aide. Depuis peu, il existe un ensemble d'outils spécifiques. | http://fits.gsfc.nasa.gov |

Le premier enseignement de cette analyse est que les principaux formats de données émanent d'organismes majeurs (NASA, NCSA, NSSDS, ISO, ..) dont le seul poids suffit à imposer un format comme standard de fait. Les volumes de données manipulés par ces organismes confèrent à ces formats, à partir du moment où ils sont mis en œuvre, des références fortes, ainsi qu'une dynamique globale leur permettant de constituer un ensemble de couches applicatives facilitant la manipulation des données.

Le deuxième enseignement est que ces formats émanent la plupart du temps des domaines scientifiques et spatiaux. Ces domaines sont en effet directement concernés par la problématique de la standardisation des données qui a été décrite dans les chapitres précédents. Les consommateurs des données produites sont rarement en relation avec les producteurs de ces données. Les scientifiques exploitant les données appartiennent la plupart du temps à des organismes (laboratoires, universités, ..) n'ayant pas de contacts avec les organisations (agences spatiales, autres laboratoires,…) les ayant produites. On trouve souvent des communautés à l'échelle mondiale, allant chercher des produits dans des catalogues distants.

Ces formats de données sont pour la plupart auto descriptifs. La description de la donnée est incluse dans la donnée elle-même, et les informations ne peuvent donc pas être intégrées et agencées librement. La création de produits respectant ces différents formats nécessite souvent une expertise ou tout au moins un effort de conception.

Les informations sémantiques associées sont souvent pauvres et associées la plupart du temps à un complément documentaire (d'où une certaine perte d'intérêt pour le format auto-descriptif).

Les couches logicielles sont élaborées et s'organisent plus ou moins de la même façon :

- des interfaces de programmation (API) de bas niveau accédant aux fichiers HDF
- des interfaces de programmation spécialisées, offrant des services de plus haut niveau à partir des services de base
- des applications directement utilisables par les utilisateurs ou les producteurs de données, bénéficiant la plupart du temps d'une interface graphique.

Ces logiciels ainsi que un ensemble de documentations relatives au projet sont disponibles sur des serveurs Internet (serveurs http ou ftp), suivant la philosophie propre aux logiciels libres.

# 7. PRESENTATION DETAILLEE DES STANDARDS DU CCSDS ET DE LEUR OUTILLAGE

Ce chapitre présente de façon détaillée les standards de description de données élaborés dans le cadre des activités du CCSDS. Ces standards permettent de couvrir les aspects sémantiques et syntaxiques de la description. Ces deux aspects complémentaires sont abordés en détail.

Les chapitres précédents ont permis d'aborder la problématique globale de la standardisation et ont offert un aperçu sur l'offre disponible. La description poussée de standards va nous permettre de mieux appréhender les apports et les possibilités d'une politique de standardisation des formats de données.

Ces standards sont, au même titre que les standards présentés ci-dessus, accompagnés d'une gamme d'outils accompagnant la donnée sur son cycle de vie. Les outils les plus représentatifs sont donc présentés.

Enfin des exemples précis de mise en œuvre nous permettront de mieux situer les apports potentiels de cette approche.

## 7.1. LE LANGAGE DE DESCRIPTION DE DONNEES EAST

Comme on l'a vu précédemment, tous les grands centres de traitement et d'archivage de données sont confrontés au problème de la pérennité et de l'intégrité des données qu'ils traitent et conservent. Ce problème présente deux aspects à savoir :

- garantir leur intégrité physique (obsolescence des supports par exemple),
- garantir la disponibilité de la description de ces données en permettant ainsi d'y accéder et de les comprendre.

Le premier point s'obtient par des techniques appropriées qui prévoient les renouvellements de support et des copies de sécurité. On trouve des produits clés en main assurant ce genre de service

Le langage EAST et la technologie développée autour de ce langage apportent une solution au deuxième point.

### 7.1.1. INTRODUCTION AU LANGAGE EAST

Le langage EAST a été conçu dans le cadre du CCSDS pour :

- autoriser une description rigoureuse et exhaustive du format de données quelconques, sans référence "volatile" à de quelconques systèmes de gestion de fichiers non pérennes,
- permettre l'accès aux valeurs de ces données grâce à des outils génériques évitant d'écrire la moindre ligne de code spécifique aux données à lire,
- permettre le formatage des données sur leur support de stockage ou d'échange en garantissant par construction leur conformité avec leur description.

Ces objectifs ont été atteints en produisant un langage de description des données axé principalement sur ses capacités de description, de génération et d'interprétation automatique, mais aussi sur sa lisibilité. Les capacités de description de EAST sont basées sur celles d'un langage existant (ADA) afin de fournir des informations complètes et non ambiguës sur le format des données traitées. L'aspect «formel » du langage permet de le rendre «compilable » et «interprétable » par des logiciels tiers afin d'accéder et de générer des données automatiquement. EAST est de plus un langage lisible, porté vers des constructions linguistiques plutôt que vers des mots clés « cryptés ». On verra cependant que l'outillage fourni permet de s'affranchir complètement de la connaissance du langage, facilitant ainsi sa mise en œuvre et son « adoption ».

Le langage EAST est aujourd'hui une norme internationale (CCSDS et ISO n° FDIS 15889). Des outils s'appuyant sur cette norme ont été développés. Ils permettent d'assister les acteurs intervenant tout au long du "cycle de vie des données".

### 7.1.2. STRUCTURE D'UNE DESCRIPTION EAST

Une description EAST d'un format de données (appelé DDR, pour Data Description Record) inclut une description syntaxique (et sémantique par certains aspects), suivie d'une description physique. La description physique rend possible l'interprétation de la série de bits rencontrés sur le média. Une description contient donc deux « packages », un pour la partie logique et un autre pour la partie physique.

La partie logique d'une description EAST comprend :

- Une description logique de tous les composants ou champs de la donnée (noms des champs, structuration, types …)
- Leurs tailles en bits
- Leur localisation dans l'ensemble des champs décrits

La partie physique d'une description EAST comprend :

- La représentation des types basiques (énumérés, entiers, réels), utilisés dans la partie logique et dépendante de la machine ayant généré les données
- L'organisation des tableaux (first-index-first ou last-index-first) utilisée par la machine ayant généré les données
- L'organisation des bits et des octets sur le média (high-order-first ou low-order-first)

La description logique précède toujours la description physique.

La séparation en deux parties distinctes permet pour une même description logique d'associer plusieurs descriptions physiques, et donc plusieurs types de machines. Par exemple un réel 32 bits sur une architecture IEEE a une description physique différente de celle d'un réel sur une architecture 1750, bien que leurs tailles en bits soient identiques. Il est à noter que les représentations utilisées pour écrire ou lire des données d'une machine particulière ne doivent pas être forcément celles de la machine support à la lecture ou à la génération (on pourra ainsi par exemple manipuler des données PC sur des machines SUN)

### 7.1.3. DESCRIPTION SUCCINTE DU LANGAGE

On a vu dans le chapitre précédent que la partie logique d'une description fournit des informations syntaxiques et dans une certaine mesure syntaxiques (informations nécessaires à l'utilisateur pour comprendre quel type de données il manipule). La partie physique fournit une description au niveau du bit qui assure la non-ambiguïté lors de l'interprétation des données.

Dans ces deux parties (ou « packages »), la syntaxe utilisée est basée sur les notions de types et d'objets. Un type est un modèle, défini une fois pour toutes, utilisé pour créer les occurrences (ou objets) du modèle.

Chaque item d'une description EAST est un objet, avec un type permettant de définire un ensemble de valeurs. Les types basiques sont les types scalaires (types numériques et énumérations décrivant des éléments simples) et les types composés (tableaux, listes, structures décrivant des séquences d'objets).

Chaque type est nommé, le nom permettant de décrire la signification du modèle (une DATE par exemple). Chaque objet est nommé, et l'on peut ainsi définir la particularité de l'occurrence (une DATE_DE_DEBUT_DE_MESURE par exemple).

Le types de base proposés sont les suivants :

- L'énuméré qui définit un ensemble ordonné d'énumérations littérales distinctes. Par exemple, un type booléen définit deux items d'énumération (VRAI et FAUX).
- Les caractères qui ont un type énuméré prédéfini.
- Les types numériques qui permettent de décrire les entiers et les réels
- Les types composés qui permettent les définitions d'objets structurés. Les types proposés par le langage sont les tableaux, les listes et les structures. Les tableaux sont des objets avec des composants indexés du même type. Le type «STRING » (chaîne de caractères) est un type de tableau prédéfini. Une liste est un ensemble ordonné de composants du même type. Une structure est un objet regroupant des composants de différents types possibles.

Une structure peut posséder des discriminants. Un discriminant permet de spécifier des structures alternatives (un ou plusieurs champs seront présents ou absents suivant les valeurs d'un autre champ) ou bien permet de définir dynamiquement la taille d'un tableau.

Le concept de typage est renforcé par le concept de sous-typage, par lequel un utilisateur peut par exemple restreindre l'ensemble des valeurs permises pour un type scalaire. Les sous types peuvent aussi être définis pour définir des index de tables.

Des clauses de représentation sont utilisées pour définir les correspondances entre les types logiques et leur représentation physique. On spécifiera par exemple que les objets d'un type donné sont représentés avec un nombre défini de bits ou que les composants d'une structure sont organisés suivant une certaine disposition.

EAST est un sur-ensemble de la partie déclarative du langage de programmation Ada. Il ne reprend pas les primitives liées à l'algorithmie. EAST étend la puissance de ce langage en décrivant non seulement les aspects logiques, mais aussi les aspects physiques.

EAST a été élaboré afin de prendre en compte les besoins les plus larges possibles d'un ensemble d'organismes du domaine spatial et ce pour différents types de domaines applicatifs (sciences, observation de la terre, …). Des données existantes (appelées données historiques) ainsi que des données futures ont été prises en compte pour la définition des besoins puis l'élaboration de ce langage. Il répond ainsi à des besoins très larges dépassant largement le cadre du domaine spatial.

Afin de faciliter la mise en œuvre de la norme EAST et d'en permettre ainsi l'accès à une large communauté d'utilisateurs, un ensemble d'outils a été développé. Ces outils, en affranchissant l'utilisateur de la difficulté inhérente à l'apprentissage d'un format ou d'un langage de description, lui permettent d'accéder à l'ensemble des avantages liés à la standardisation des données.

## 7.2. LE LANGAGE DEDSL POUR LES DICTIONNAIRES DE DONNEES

Lorsque l'on décrit des données on peut se placer à plusieurs niveaux :

- au niveau inventaire (existence d'un jeu de données)
- au niveau syntaxe (description des données sur leur support physique)
- au niveau sémantique (description de la signification des données)

Les inventaires sont en général traités sous la forme de catalogues. La syntaxe quant à elle est décrite par des documents de description de fichiers plus ou moins formels. Comme on l'a vu dans les chapitres précédents, un produit composé de données peut être généré ou distribué avec un standard de format (FITS, CDF, HDF) ou de description (EAST). Cette information, principalement syntaxique, peut ne pas être facile à comprendre et un niveau de description sémantique est alors nécessaire. Cette description va donc mener à la définition d'un dictionnaire d'entités. Le langage DEDSL permet de traiter la description sémantique et donc de mettre en œuvre les dictionnaires de données.

### 7.2.1. LES DICTIONNAIRES DE DONNEES

Un dictionnaire est un mécanisme capable d'organiser un ensemble d'informations de façon cohérente et compréhensible car il est avant tout destiné aux utilisateurs humains. Il leur permet d'accéder à la signification ainsi qu'à d'autres informations utilisées dans la définition et la génération de données.

Les dictionnaires de données peuvent couvrir deux grands types de besoin :

- Description d'un domaine : on trouvera dans ce type de dictionnaire la définition de tous les concepts utiles pour modéliser le domaine. Par exemple on peut imaginer le dictionnaire de l'observation de la terre ou bien celui de la physique nucléaire. Ce type de dictionnaire a pour but de faciliter la compréhension entre membres d'une même communauté d'intérêt.
- Description d'un produit : on trouve dans ce genre de dictionnaire la description des divers champs dont un produit (en général échangé sous forme de fichiers) est composé. Pour chaque champ ou type de champ on disposera de la liste des attributs renseignés qui le décrivent.

Un dictionnaire peut être décrit de plusieurs façons. Il peut être défini en langage naturel, sous la forme d'un ensemble de paragraphes constituant un document accompagnant le produit. Il peut être aussi décrit par des attributs dont le type et les modes de définitions correspondent à des formats ou des langages particuliers.

Les individus ou organisations ayant à recevoir et comprendre une grande variété de produits, peuvent passer un temps important à la compréhension de ces dictionnaires. Si ces dictionnaires ne répondent pas à un format ou une norme particulière, il sera alors difficile de mettre en œuvre des outils les assistant dans la présentation et la compréhension des données. Ainsi la mise en œuvre des dictionnaires dans le cadre d'organisations nécessite au préalable la définition de concepts standards utilisés pour la constitution de ce patrimoine. Pour accompagner ces concepts d'outils génériques il sera donc nécessaire de définir une représentation standard associée.

Le langage DEDSL définit un ensemble de concepts sous la forme d'attributs, suffisamment généraux pour être appliqués de façon assez large. Il fournit aussi une représentation pour la définition de nouveaux attributs.

### 7.2.2. INTRODUCTION AU LANGAGE DEDSL

Le langage DEDSL a été conçu dans le cadre du CCSDS avec comme objectif la prise en compte de l'existant ISO dans le domaine. Il spécifie un certain nombre de mots clés auxquels on peut affecter une valeur pour décrire des données.

Le langage définit les concepts de nom, de signification, d'unité ainsi qu'un ensemble d'autres attributs pouvant être utilisés communément dans les dictionnaires. La façon dont ces mots clés sont spécifiés est elle-même normalisée de façon à ce que les utilisateurs puissent définir des mots clés supplémentaires qui leur seraient utiles. Une méthode permet donc, en complément, d'étendre cet ensemble d'attributs standards. Etant donné que ce mode d'élaboration de dictionnaires a pour objectif d'être utilisé dans tous les domaines, seulement un nombre limité de ces attributs sont obligatoires.

Les attributs proposés n'offrent pas les moyens de décrire les relations entre les entités. Ils ne permettent pas non plus de décrire les représentations physiques des entités. Le langage est donc complémentaire aux formats et langages de description des données tels que EAST.

L'ensemble des mots clés et des valeurs qui leur sont affectées par des utilisateurs décrivant des données forment des dictionnaires de données.

La norme en est à son avant-dernier stade («livre rouge») ce qui signifie prêt pour une dernière revue par les agences spatiales avant son passage à l'état définitif applicable.

A partir de tels dictionnaires normalisés on peut produire de la documentation dans des formats «projets».

**Exemples de description**

On fait figurer ci-après quelques attributs standards du langage DEDSL utilisés dans des descriptions de données :

```
BEGIN_GROUP = ENTITY_DEFINITION;

NAME = SATELLITE_ID ;

CLASS = DATA_FIELD ;

DEFINITION = "Identification of the satellite " ;

INHERITS_FROM = A_SATELLITE_ID ;

COMPONENT = FAMILY_NAME ;

COMPONENT = NUMBER ;

END_GROUP = ENTITY_DEFINITION;
```

Cette description concerne un champ identifiant un satellite. L'identification consiste en deux champs donnant l'un la famille (ex : SPOT) et l'autre le numéro dans la famille. Ce champ est de type "a_satellite_id", type lui-même décrit par ailleurs. Il hérite donc de tous les attributs liés à ce type en particulier pour son codage physique (ASCII, binaire, longueur etc...).

### 7.2.3. APPORTS

Cette façon de traiter la description de donner présente plusieurs avantages :

- elle induit grâce aux dictionnaires de domaines une normalisation de fait de la description des produits,
- elle ouvre la porte à une gestion séparée du fond et de la forme des descriptions. Avec une même syntaxe canonique en DEDSL on peut concevoir des mises en formes différentes pour chaque contexte d'utilisation (Spatial, Défense, Energie, Science etc...).
- la normalisation des mots clés autorise la réalisation d'outils de recherche génériques capables de travailler sur les données de projets, de domaines et d'organismes différents.

Ainsi, le langage DEDSL pourra être utilisé :

- par les producteurs de données, pour la construction de dictionnaires décrivant de façon formelle les entités du produit,
- par les utilisateurs des données, en offrant la capacité à comprendre les données reçues,
- par des organisations ou organismes souhaitant homogénéiser les attributs à renseigner et à utiliser pour les entités manipulées,
- par une communauté (sciences, archives, …) souhaitant établir une standardisation du contenu des dictionnaires du domaine, pouvant à terme être utile à l'élaboration d'un dictionnaire général,
- par les organisations ou communauté souhaitant échanger les contenus des dictionnaires de données de façon standardisée ou bien faciliter l'interopérabilité entre les dictionnaires.

## 7.3. LES OUTILS DE LA TECHNOLOGIE EAST

Les données ont un cycle de vie. Par cette affirmation on entend qu'elles sont imaginées, conçues, décrites, produites et enfin consommées. Cet ordre n'est peut-être pas une évidence pour qui a eu à décrire des données déjà bien existantes sur leur support physique afin de pouvoir les consommer.

Pourtant, le langage EAST et les outils associés permettent de procéder en respectant un tel cycle.

- l'outil OASIS permet de produire des descriptifs de la syntaxe des données (description des bits sur leur support et de l'interprétation qu'il faut en faire pour obtenir des valeurs significatives dans les applications qui les consomment), ainsi que de la sémantique. L'utilisateur produit du EAST sans avoir à connaître la syntaxe de ce langage,

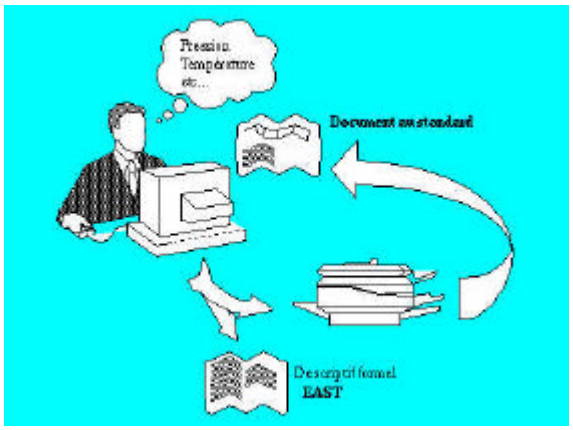


**fig. 6 : Synoptique de l'utilisation d'OASIS**

- un générateur permet ensuite (en s'appuyant sur le descriptif en langage EAST créé avec OASIS) de formater les données en leur assurant une conformité "congénitale" avec leur description,

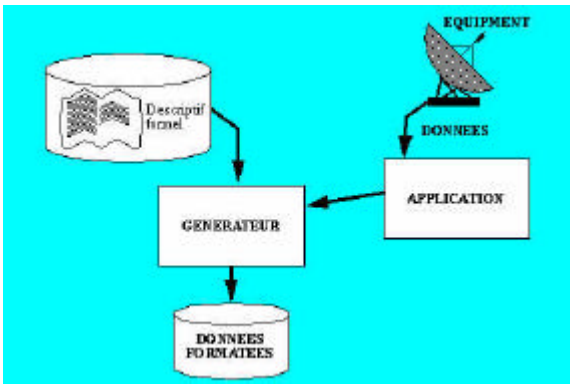


**fig. 7 : Synoptique de l'utilisation du générateur**

- un outil d'accès (l'interpréteur) permet la lecture des données en utilisant lui aussi le descriptif en langage EAST,

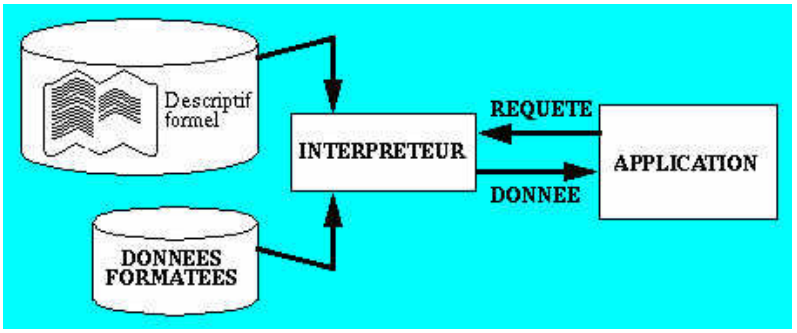


**fig. 8 : Synoptique de l'utilisation de l'interpréteur**

- enfin, des outils de visualisation s'appuient sur l'interpréteur pour présenter les données aux utilisateurs.

## 7.4. PRESENTATION DETAILLEE DES OUTILS

Ce chapitre présente de façon détaillée les fonctionnalités et la mise en œuvre des principaux outils de la technologie.

### 7.4.1. L'OUTIL OASIS

#### *7.4.1.1. GENERALITES*

L'outil OASIS se présente sous la forme d'une interface homme/machine graphique adaptée à la saisie des caractéristiques de données.

Il permet de créer et de maintenir aisément la description des données. Si des évolutions sont nécessaires, on part de l'état courant de la description et on apporte les modifications. On peut alors immédiatement produire le nouveau descriptif en langage EAST incluant les évolutions.

L'outil OASIS qui servait initialement à décrire la syntaxe des données (leur format sur leur support) a été complété pour permettre la saisie de la sémantique des données. Il permet la configuration des attributs nécessaires à une telle description et la saisie des valeurs d'attributs pour chaque concept ou champ.

Il permet bien évidemment la production d'un dictionnaire normalisé en langage DEDSL.

Cette normalisation permet la réalisation de post-processeurs permettant l'édition de tels dictionnaires sous la forme de documents plus ou moins traditionnels (word, postscript, html ou framemaker...).

Ainsi, pour les produits de données on associe dans une même gestion de la description (par OASIS) la description syntaxique et sémantique. Par la suite, les évolutions sont plus faciles à gérer qu'avec des outils de production documentaire généralistes.

Il est disponible sur station Unix SUN. Une version PC est à l'étude.

#### *7.4.1.2. MODES D'UTILISATION*

OASIS permet au concepteur de la donnée de la définir et de la décrire sous la forme d'un arbre dont les nœuds sont les niveaux structurants (structures, tableaux, listes) et dont les feuilles sont des données scalaires (numériques, textuelles, énumérées). Suivant les objectifs des utilisateurs, l'outil peut supporter une approche purement sémantique (correspondant à la phase amont de conception de la donnée) ou bien purement syntaxique (correspondant à la phase de description de la donnée). Des approches mixtes sont aussi possibles. Ainsi, les itérations entre les concepteurs et descripteurs de la donnée peuvent être basées sur le même outil, les auteurs manipulant une seule description, suivant plusieurs points de vue.

Au niveau de l'approche syntaxique, l'utilisateur décrit les caractéristiques formelles utilisées lors de la phase d'interprétation : nom du champ, taille, position, valeurs possibles (limites pour les numériques, valeurs possibles pour les énumérés). Les notions de dicriminants présentées sur le chapitre relatif à la norme EAST sont prises en compte. Les utilisateurs peuvent décrire leurs propres types afin d'assurer une cohérence dans un même descriptif, puis utiliser ces types dans d'autres descriptifs afin d'assurer une cohérence globale. Quand les types de bases sont utilisés, OASIS propose un ensemble de valeurs par défaut (la taille en bits du champ par exemple)

afin d'en accélérer la saisie. Des contrôles de validité sont effectués à chaque action de l'utilisateur (empêchant par exemple l'utilisation de deux champs du même nom dans une structure).

Au niveau de l'approche sémantique, OASIS associe à chaque champ un ensemble d'attributs obligatoires ou optionnels que les concepteurs de la donnée ont à renseigner.

Des contrôleurs de cohérence syntaxiques et sémantiques vérifient que le descriptif est dans un état suffisamment avancé et cohérent pour générer les descripteurs EAST et DEDSL.

La figure suivante présente un arbre OASIS avec la sélection d'un champ énuméré. La fenêtre «information» contient les caractéristiques syntaxiques du champ sélectionné. La fenêtre « Implantation » représente les tailles et les positions des champs de la structure «PARAMETERS ». Cette vue particulière offre une représentation à plat de la structure. Chaque champ de la structure apparaît comme une boîte avec son nom, sa taille, son bit (ou octet) de début et son bit (ou octet) de fin. Les champs optionnels (discriminés) sont représentés avec une couleur spécifique et sont placés sous la ligne de base en autant de lignes que de combinaisons possibles.

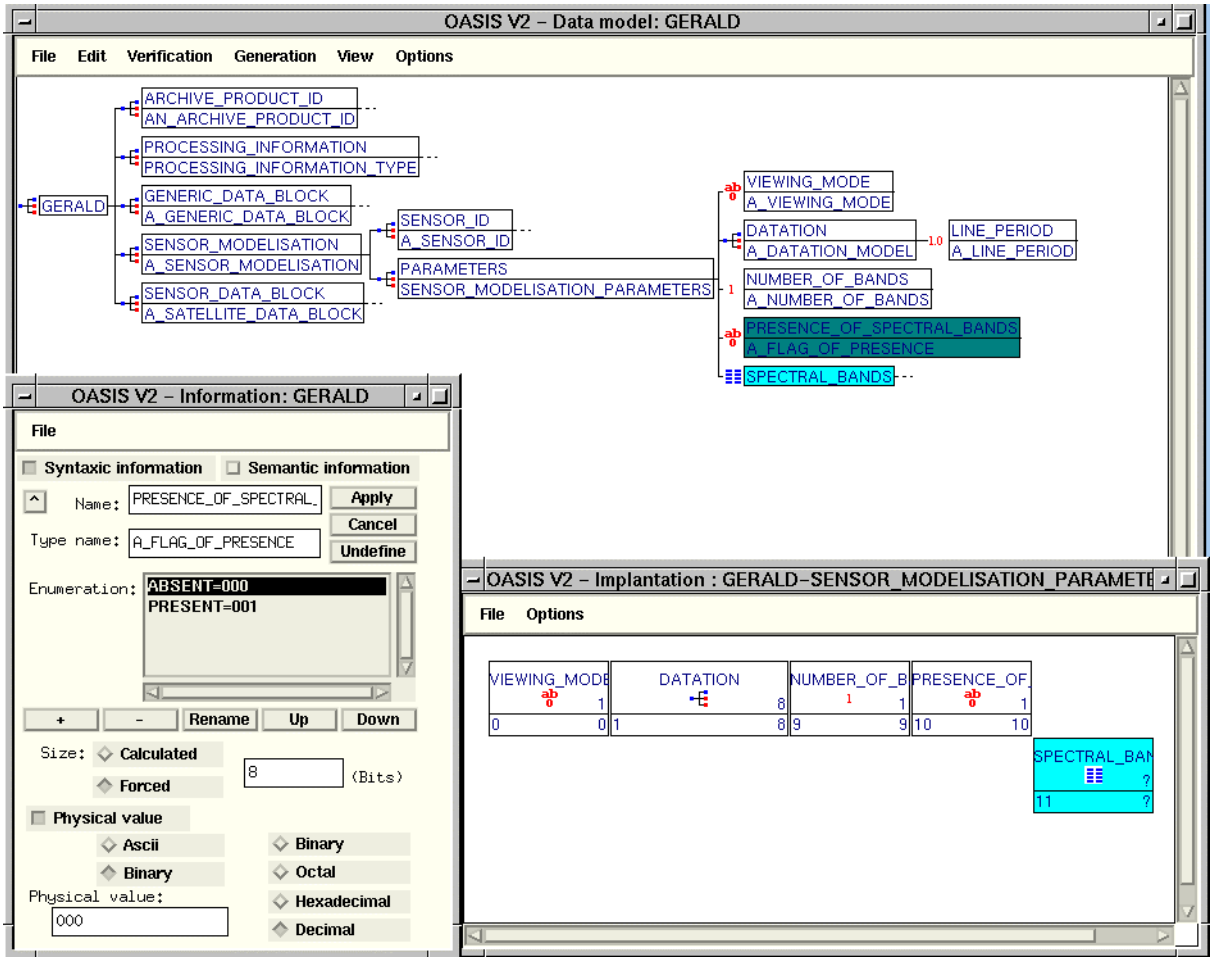


**fig. 9 : Vue générale de l'interface utilisateur OASIS**

OASIS peut analyser et importer des descriptifs EAST ainsi que des dictionnaires écrits en DEDSL. On peut ainsi récupérer des dictionnaires existants afin d'intégrer les informations disponibles dans les descriptifs.

OASIS permet de générer :

- Une description syntaxique au format EAST
- Un dictionnaire du produit au format DEDSL
- Une documentation aux formats Word, FrameMaker ou Postscript, incluant les arbres, les descriptions des structures ainsi qu'un ensemble de champs. La liste des éléments présentés dans la documentation est configurable.

De nouvelles fonctionnalités sont en cours de développement pour étendre les formats disponibles (XML, HTML, …).

### 7.4.2. L'INTERPRETEUR DE DONNEES

L'interpréteur de données EAST est un paquetage ADA, offrant des primitives qui permettent à une application "utilisateur" de lire des fichiers de données décrites au format EAST.

L'interpréteur peut lire des fichiers présents sur disque dur, ou des données présentes en mémoire, et interpréter des données générées sur SUN, PC, VAX (la liste étant non exhaustive).

L'interpréteur est piloté par une application via des requêtes.

Bien qu'étant écrit en ADA, l'interpréteur peut être utilisé depuis une application ADA (avec une référence par clause "with" à l'interpréteur), mais aussi depuis une application autre (actuellement C et Fortran 77) car il possède les interfaces programmatiques adéquates.

L'interpréteur se présente sous la forme d'une bibliothèque de fonctions. Cette bibliothèque est utilisable depuis les applications clientes. Elle fournit un ensemble de fonctionnalités parmi lesquelles on peut citer :

| | | |
|---|---|---|
| select_DDR | : | Cette opération permet de sélectionner un fichier descriptif de données, et de générer sa forme intermédiaire associée qui est conservée en mémoire (tout au long de l'exécution de l'application) et sur fichier. |
| load_next_record_from_file | : | Cette opération permet de charger en mémoire une variable de haut niveau (appelée aussi record) à partir d'un fichier. |
| load_next_record_from_mémory | : | Cette opération permet de charger en mémoire une variable de haut niveau à partir d'une zone mémoire. |
| get_number_elements_of_array | : | Cette opération donne le nombre d'éléments d'une donnée typée tableau (à partir de son chemin d'accès EAST) statique ou dynamique. |
| get_data_entity | : | Cette opération permet de récupérer et d'interpréter une occurrence d'une donnée (à partir d'un chemin d'accès).. |

Les services fournis par l'interpréteur doivent être enchaînés dans un ordre défini. L'application initialise l'interpréteur ("start"), puis, pour chaque groupe de données à interpréter, il faut analyser son descriptif ("select_DDR"), charger un bloc de données correspondant au descriptif (ex :"load_next_data_block_from_file"), puis récupérer une ou plusieurs occurrences des entités voulues dans ce bloc ("get_data_entity_ascii").

La bibliothèque permet donc de réduire de façon importante les volumes de code en diminuant ainsi les coûts de développement et de maintenance. De plus, une évolution du format de données entraînant des modifications du fichier, n'impliquera pas dans la plupart des cas une modification des codes correspondants. Seul le descriptif (modifiable avec OASIS) devra être mis à jour.

Un utilitaire « ascii_dump », développé avec cette bibliothèque, est délivré en standard avec l'interpréteur. Celui ci prend en entrée un fichier de données et son descriptif EAST associé puis génère un fichier résultat où les données sont présentées champ par champ au format ascii. L'utilitaire permet de présenter en particulier les données posant des problèmes (comme par exemple un entier hors bornes).

### 7.4.3. LE GENERATEUR DE DONNEES

Le générateur de données EAST est un paquetage ADA, offrant des primitives qui permettent à une application "utilisateur" de créer des données dans un format décrit avec le langage EAST.

Les données générées sont, soit créées aléatoirement, soit converties à partir de valeurs indiquées par l'utilisateur. L'outil permet également de lire des données décrites par le descriptif EAST sélectionné et de modifier certaines d'entre elles de façon aléatoire ou à l'aide de valeurs imposées par l'utilisateur avant de les écrire en mémoire ou sur support.

Le générateur est piloté par une application utilisatrice via des requêtes.

Bien qu'étant écrit en ADA, il peut être utilisé depuis une application ADA (avec une référence par clause "with" au générateur), mais aussi depuis une application autre (C ou tout autre langage).

Le générateur se présente sous la forme d'une bibliothèque de fonctions. Cette bibliothèque est utilisable depuis les applications clientes. Elle fournit un ensemble de fonctionnalités parmi lesquelles on peut citer :

| | |
|---|---|
| select_DDR : | Cette opération permet de sélectionner un fichier descriptif de données, et de générer sa forme intermédiaire associée (qui est conservée tout au long de l'exécution de l'application). |
| Read : | Cette opération permet de lire un bloc de données de référence correspondant au descriptif EAST précédemment sélectionné à partir d'un fichier. |
| set_data_entity #1 | Cette opération permet d'affecter des valeurs imposées aux données élémentaires (entier, réel, énumération, chaîne de caractères). Par défaut, le générateur contrôle la validité de ces valeurs mais l'utilisateur a la possibilité de supprimer explicitement ce contrôle. |
| set_data_entity#2 : | Cette opération permet de créer des données de façon aléatoire. Cette création peut porter sur une donnée complexe (record au tableau) comme sur une donnée élémentaire. |
| write | Cette opération permet d'écrire un bloc dans un fichier une fois que toutes les données élémentaires ont une valeur associée. |
| verify : | Cette opération permet de savoir si la valeur choisie pour l'affectation est correcte pour la donnée choisie (chemin EAST) d'après le type qui lui est associé. |

Les services fournis par le générateur doivent être enchaînés dans un ordre défini. L'application initialise le générateur ("start"), puis sélectionne le descriptif EAST correspondant au format des données à générer ("select_DDR) et l'analyse. L'application peut ensuite charger un bloc de données de référence correspondant au descriptif EAST précédemment sélectionné (étape optionnelle). Elle crée alors les données au format EAST sélectionné (si la génération ne s'effectue pas à partir d'un bloc existant) ou elle demande la modification des données déjà chargées au préalable. La création et la modification s'effectuent suivant deux modes :

- soit de façon aléatoire
- soit par affectation de valeurs imposées exprimées sous la forme de chaînes ASCII
- soit par affectation de valeurs binaires

Une fois toutes les données renseignées, l'écriture d'un bloc est possible.

### 7.4.4. LE GENERATEUR MODIFIEUR DE DONNEES (DUW)

L'IHM pour la mise à jour et la consultation de données, appelé DUW (pour Data Update Wizard) propose une interface conviviale pour générer des fichiers de données aux normes de la technologie EAST, que ce soit à partir de données existantes ou non, ainsi que pour la consultation simple des données.

Globalement, les fonctionnalités sont les suivantes :

- Sélectionner un fichier de description qui répond aux normes de la technologie EAST, pour visualiser à l'écran les données sous une forme d'arbre,
- Sélectionner un type d'opération (consultation / génération),
- Sélectionner un mode pour l'opération précédente,
- Sélectionner le cas échéant des fichiers source et des fichiers de résultat,
- Consulter et/ou générer des fichiers de données.

Cette IHM peut être utilisée à partir d'une station Sun Solaris 2.7 et à partir d'un terminal Vax OpenVMS 7.2. Développée en java, elle devrait bientôt être disponible sur plate-forme PC.

L'outil proposé offre un ensemble de services étendus avec en particulier :

- La navigation dans les graphes des descriptifs (même type de service que celui proposé à OASIS mais tourné vers l'utilisateur),
- La navigation dans l'arbre des données en cours de lecture ou de génération,
- La génération de données « from scratch » ou les valeurs des champs seront renseignées de façon manuelle ou aléatoire,

- La génération de données à partir de fichiers d'entrée, avec la possibilité de modifier des champs (mode Update),
- L'accès en ligne au contenu du dictionnaire des données,
- La gestion des discriminants et des champs optionnels.

Ainsi l'ensemble des fonctionnalités proposées par les outils «Interpréteur » et «Générateur » sont accessibles par cette interface à tous les utilisateurs.

La figure suivante donne un aperçu de l'environnement fourni par cet outil :

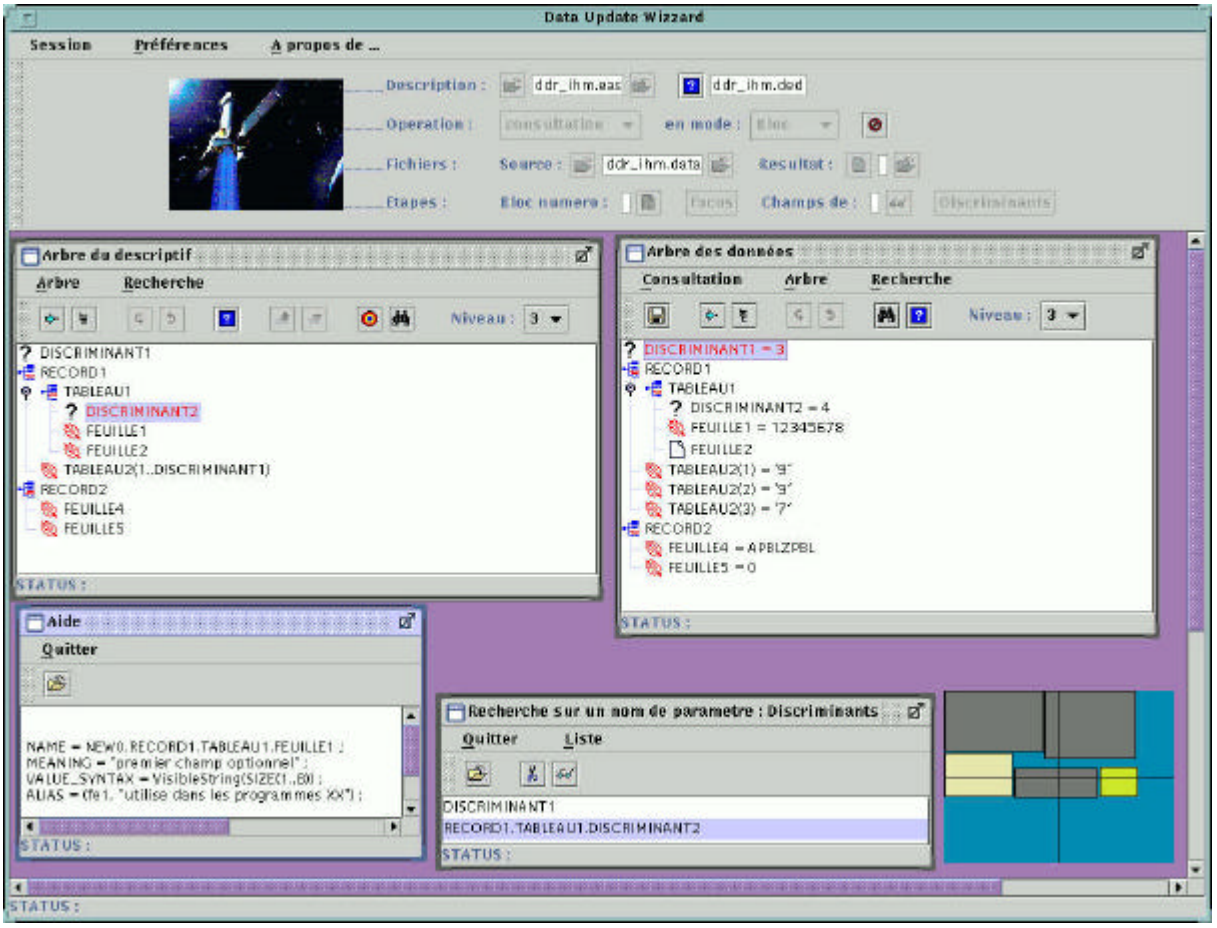


**fig. 10 : Vue générale de l'interface du DUW**

Dans cette même fenêtre l'utilisateur a accès à la plupart des fonctions fournies par l'outil :

- Le panneau de contrôle situé en haut de la fenêtre, permettant la sélection des descriptifs, des données, ainsi que l'activation des principales fonctionnalités,
- L'arbre du descriptif présentant la structure et le type des données,

- L'arbre des données permettant de consulter, générer ou modifier le contenu des fichiers de données sélectionnés,
- La fenêtre d'aide offrant les informations sémantiques associées aux champs,
- La fenêtre de définition des discriminants (en mode génération).

### 7.4.5. L'EXTRACTEUR DE DONNEES (DEW)

L'IHM de l'extracteur de données appelé DEW (pour Data Extraction Wizard) propose une interface conviviale pour la sélection de paramètres d'une extraction de données. Elle permet à un utilisateur de visualiser des données dont la structure est celle d'un arbre. Ces données doivent avoir été au préalable stockées dans un fichier de description répondant aux normes de la technologie EAST. A partir de cet arbre, l'utilisateur peut sélectionner des données pour réaliser une extraction.

Globalement, les finalités sont les suivantes :

- Sélectionner un fichier de description qui répond aux normes de la technologie EAST, pour visualiser à l'écran les données sous une forme d'arbre,
- Sélectionner dans cet arbre les données dont on souhaite connaître la valeur,
- Pour chacune de ces données, définir :
  - Le format, c'est-à-dire le nombre de caractères ASCII sur lesquels on veut que le résultat soit écrit,
  - Si la donnée contient un tableau, l'intervalle des valeurs à extraire,
  - Si les données sont associées à des dates, la plage de dates recherchée,
  - La position d'une donnée dans la liste des données.

Cette IHM peut être utilisée à partir d'une station Sun Solaris 2.7 et à partir d'un terminal Vax OpenVMS 7.2. Elle se présente de façon similaire à celle du DUW.

L'utilisation de cet outil est particulièrement bien adaptée à l'extraction de données en vue de leur exploitation sur des logiciels de visualisation graphiques du type PV-Wave.

## 7.5. APPORTS DES OUTILS

Les descriptions de données produites grâce à OASIS assurent la pérennité de ces données. En effet, EAST ne supporte pas les non-dits implicites d'une description informelle. On ne pourra pas, par exemple, se contenter de dire qu'une donnée est codée sous la forme d'un entier. Il faudra préciser dans un premier temps si le codage est binaire ou ascii et en fonction de ce premier choix toutes les caractéristiques plus précises de ce codage.

Les applications utilisant les outils EAST pour leurs lectures et écritures sont aisément maintenables. En effet, tout changement de format n'entraînant pas de changement de comportement est transparent pour l'application que l'on n'a pas besoin de retoucher. Seul le descriptif EAST devra être modifié pour prendre en compte les évolutions. Pour peu que l'on prenne la précaution d'inclure dans les données un champ donnant la version de ces données on peut avoir des descriptifs compatibles ascendants qui permettent de traiter le nouveau format tout en restant capable de traiter l'ancien.

La rigueur des descriptifs supprime quasiment tous les problèmes d'interface entre les logiciels qui s'échangent des données.

On peut, grâce au générateur, simuler des données au bon format (de façon contrôlée et/ou aléatoire) avant que les processus destinés à les produire n'existent.

Les outils de base (« Interpréteur » et « Générateur ») sont utilisables pour qui prend la peine de développer un applicatif client. Les interfaces de consultation, génération et extraction apportent le complément essentiel permettant aux utilisateurs de disposer de ces utilitaires immédiatement, sans développer une ligne de code.

Cette offre se décline donc en différentes « couches », des outils de base (type interface de programmation) aux outils graphiques, que nous pouvons modéliser grâce au schéma suivant :

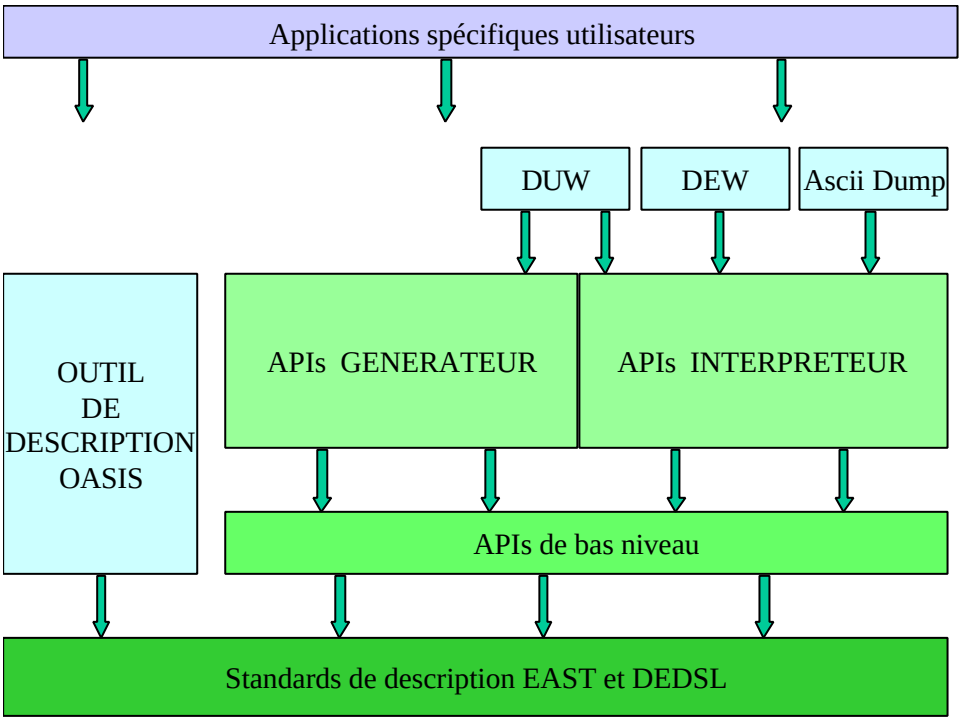


**fig. 11 : Les niveaux d'interaction avec les standards CCSDS**

## 7.6. DIFFUSION

La disponibilité de la bibliothèque associée au langage est indispensable à la pénétration de la norme EAST dans le domaine de la description des données. Une forte demande existe, notamment au niveau de l'ensemble des agences spatiales.

Tous les outils de la technologie EAST et leurs documentations associées sont donc en cours d'installation sur le serveur de logiciels libres du CNES pour leur assurer la plus large diffusion. Ils sont maintenus par le CNES et CS SI qui traitent les éventuelles anomalies et les propositions de modification.

Il existe également un serveur Internet présentant cette technologie (adresse en cours de finalisation). Ce serveur pointera sur le serveur de logiciel libre dès son ouverture au public.

On peut trouver les normes et standards CCSDS/ISO sur le serveur : http://ccsds.org/p2/home.html

## 7.7. REFERENCES D'UTILISATION DE LA TECHNOLOGIE EAST/DEDSL

### 7.7.1. LES REFERENCES PRINCIPALES DE LA TECHNOLOGIE

Le langage EAST a été utilisé pour les données de nombreux projets, principalement liés au domaine spatial. Cette utilisation opérationnelle connaît une très forte croissance depuis trois ans. Auparavant, la technologie (standards, normes et outils) était en cours de développement. Parmi ces projets, on peut citer :

SPOT : pour la description du format d'archive des images (format GERALD pour SPOT 1,2,3,4 en cours d'extension SPOT5)

CDPP : Centre de Données de la Physique des Plasmas, diffusant les données des satellites et expériences Arcad3, Phobos, Vicking, Interball, Cluster.. L'adresse Internet du serveur est http://cdpp.cesr.fr. Les produits sont associés à leur descriptif EAST au sein de l'archive. Le CDPP propose un service de transformations génériques des produits avant la livraison aux utilisateurs. Il se base sur l'utilisation de ces descriptifs EAST à travers l'outil d'interprétation des données.

SSALTO : Segment Sol des satellites (Jason, Doris, ENVISAT...). Ce segment sol utilise les descriptions des télémesures et des produits au format EAST et intègre les outils de lecture et écriture. L'ensemble des interfaces entre les différents sous-systèmes est défini en EAST.

AMS-2 : Archive Management System (V2) Projet ESA (ESRIN) avec Matra (MS&I) et CS SI. Le système associe aux données archivées des descriptifs EAST, et gère donc les descriptions de données et de métadonnées d'observation de la terre (ex : Données du satellite Landsat). L'outillage EAST est utilisé pour l'extraction générique de données permettant aux services de l'archive de définir et générer des sous produits des produits d'archive.

Hélios II : Comme dans le cadre du projet ENVISAT/SSALTO, le projet Hélios à choisi de décrire avec EAST l'ensemble des interfaces entre les différentes composantes du segment sol. Le projet fait une utilisation intensive des outils EAST de lecture et écriture ainsi que de l'outil OASIS.

L'ensemble des descriptifs EAST intégrés dans ces différents projets sont produits avec OASIS.

Il est à noter que dans un premier temps les projets intègrent la norme EAST pour venir ensuite à l'intégration du standard DEDSL. Les dictionnaires de ces différents domaines sont en cours d'élaboration. De ce fait l'utilisation de la sémantique ne connaît pas un stade aussi avancé que celle qui est faite de la syntaxique.

### 7.7.2. LES CHAMPS D'APPLICATION

Comme on le voit avec les références présentées ci-dessus, les champs d'application de la technologie EAST sont variés.

Au niveau des systèmes d'archive l'introduction des standards de description des données permet tout d'abord de pérenniser l'information descriptive. Elle permet de plus d'offrir un ensemble de services à valeur ajoutée tels que la possibilité de transformer de façon générique les données avant leur livraison, ou bien encore d'offrir une documentation en ligne sur les caractéristiques syntaxiques et sémantiques de ces données. Les informations distribuées aux utilisateurs sont fiables et non ambiguës.

Les applications manipulant les données (centres de traitement de données scientifiques, centres de contrôle des satellites, …), en utilisant la technologie, réduisent les efforts de développement et de maintenance en s'assurant de l'adéquation congénitale des données à leur description.

Les systèmes informatiques complexes sont souvent composés de grands sous-ensembles (appelés sous-systèmes) pour lesquels il faut définir l'ensemble des interfaces internes. Ces sous systèmes sont souvent développés en parallèle. Une des difficultés principales dans ce type de mise en œuvre provient du fait que les interfaces entre ces sous-systèmes amènent souvent des problèmes dus au manque de rigueur dans leur définition et donc à la part importante laissée aux possibilités d'interprétation par les différentes équipes de développement (pouvant appartenir à des sociétés ou départements distincts). Les conséquences lors des phases d'intégration sont souvent coûteuses (retours arrières sur la phase de codage).

Dans ce contexte particulier l'utilisation de la technologie EAST pour la définition des interfaces entre les sous systèmes permettra :

- D'éviter les ambiguïtés dans la définition des interfaces,
- De documenter ces interfaces,
- De garantir grâce à l'utilisation du générateur la conformité des fichiers d'interface fournis par un sous système,
- D'éviter grâce à l'interpréteur de données, les erreurs de lecture des fichiers d'interface fournis,
- De produire grâce au générateur, et en avance de phase dans le cycle de vie des développements, des fichiers d'interface fictifs permettant de tester séparément les cas nominaux et dégradés de fonctionnement des sous-systèmes.

La standardisation des données se révèle donc être un outil indispensable pour les grands maîtres d'œuvres souhaitant maîtriser et réussir les développements de systèmes complexes.

# 8. CONCLUSION GENERALE SUR LES FORMATS ET LES STANDARDS DE DESCRIPTION

**Bilan sur l'offre des standards de format et des standards de description**

Aujourd'hui une promotion très active de certains formats de description de données tels que CDF et HDF est effectuée. Ces formats sont, comme on l'a vu, largement diffusés. Ils sont de plus accompagnés d'une gamme d'outils permettant des manipulations poussées. Cette approche permet d'offrir une solution cohérente, outillée, répondant à un besoin important. En conclusion, ce type de formats semble particulièrement bien adapté à un contexte de données scientifiques ou les utilisateurs demandent un accès facile et adapté aux produits (par des outils de visualisation par exemple) et cherchent le bénéfice d'une dynamique globale sur le format.

Cependant, ces types de formats sont coercitifs. Ils contraignent le concepteur de la donnée à respecter le format choisi. S'ils sont particulièrement bien adaptés à certaines données scientifiques, ils ne sont pas « généralistes » et s'intègreront donc difficilement dans une politique volontariste de standardisation des données, que ce soit dans le cadre d'un projet, d'un département ou bien d'une organisation. Ainsi, la prise en compte des données historiques (données existants avant l'arrivée du format) est impossible, à moins de se lancer dans une importante campagne de «réhabilitation » des données. De plus le patrimoine de l'organisation est fragilisé, en le liant à l'existence et la pérennité d'un format particulier.

Les formats CDF et HDF sont développés et promus par la NASA qui a abandonné ses propres travaux dans le domaine des données à format libre. EAST est aujourd'hui le seul standard et seule norme de description de données à format libre. Son statut de norme ISO lui assure stabilité et pérennité.

EAST offre des avantages équivalents à ceux des formats intrusifs en fournissant des bibliothèques et des outils de manipulation des données. Il permet de plus de prendre en charge les données historiques, d'assurer la pérennité des données et donc de pouvoir s'inscrire dans une réelle stratégie de standardisation des données. Une organisation n'est pas « liée » à EAST comme elle sera liée à un format tel que HDF.

Les liens de EAST avec le langage DEDSL permettent d'aborder de façon pertinente et complète la question de l'élaboration des dictionnaires de données.

La question des liens possibles entre des formats tels que HDF et la norme de description EAST est souvent évoquée. En fait HDF et EAST ne sont pas, comme on l'a vu dans les paragraphes précédents, placés sur le même terrain. Ils ne s'inscrivent pas dans les mêmes objectifs, dans la même stratégie. En cela le passage d'une donnée décrite en EAST vers un format tel que HDF (ou vice-versa) a un intérêt assez limité. Néanmoins EAST peut arriver en complément des «trous » HDF, afin par exemple de définir de façon précise la structure et le contenu d'un tableau HDF ou bien grâce au langage DEDSL d'apporter un complément sémantique.

**Bilan sur les enjeux de la standardisation**

La standardisation des formats de données peut s'insérer dans une politique globale de gestion des produits et de sécurisation des systèmes d'information. Elle offre un ensemble de principes et de solutions pouvant implémenter ou compléter de façon pertinente les mécanismes mis en place.

Ainsi, les fichiers en entrée des systèmes peuvent être analysés en profondeur pour vérifier leur conformité au format attendu. Une fois ce contrôle effectué, on peut alors garantir aux systèmes consommateurs la qualité des données manipulées.

Un des moyens les plus sûrs pour assurer la conformité des produits livrés aux formats attendus, est l'utilisation en amont des outillages de génération des données. Cet outillage pourra de plus être utilisé avant la mise en exploitation des systèmes. En effet, la génération de données d'entrée du système permettra de tester ses comportements lors des réceptions de données nominales ou dégradées.

La structuration des données par les formats peut être de même utilisée pour définir des « vues » sur les produits, limiter l'accès à des champs restreints. Ces « vues » restreintes peuvent de même servir de base à la définition et l'extraction de sous produits d'archive adaptés aux consommateurs.

Comme on le voit, les champs d'application de la standardisation de données sont nombreux et variés. Les standards et outils peuvent être mis en œuvre des phases de spécification et de développement des systèmes à la phase d'exploitation. Le choix entre un standard de format ou un standard de description est un choix fondamental. Il devra être adapté à la stratégie globale des projets et des organismes, en fonction de leurs environnements, de l'existant, et des objectifs.