La perception du problème des métadonnées CNES

Paul Kopp

Centre National d'Etudes Spatiales (CNES)

Paul.Kopp@cnes.Fr

Le besoin de lier ce qui est dispersé

« Toutes choses étant causées et causantes, aidées et aidantes, médiates et immédiates, et toutes s'entretenant par un lien naturel et insensible qui lie les plus éloignées et les plus différentes, je tiens impossible de connaître les parties sans connaître le tout, non plus que de connaître le tout sans connaître particulièrement les parties. »

(Pascal, *Pensées*)

La vision traditionnelle

- Un souci constant du CNES :
  - préserver les données issues de ses missions spatiales
  - les restituer dans les meilleures conditions de confort possible
  - au bénéfice des communautés utilisatrices
- Pour cela, mise en place systématique d 'un « segment sol »
  - de traitement,
  - d 'archivage
  - de diffusion de données

  développé en même temps que le « segment spatial »

  (segment spatial + segment sol = système spatial réalisé pour les besoins de la mission)

La vision traditionnelle remise en cause

- Extension de l 'horizon de l 'archivage
  - de quelques années (autrefois) à plusieurs dizaines d 'années (aujourd'hui)
    - émergence de séries temporelles longues (ex. climatologie)
    - unicité de certaines observations (et des données résultantes)
  - nécessité d 'organiser les archives de données en conservatoire actif d 'un patrimoine réutilisable
- Développements conduits en coopération
  - toutes les données ne sont plus au CNES
  - les utilisateurs des données vont chercher ailleurs, en d 'autres archives, les compléments nécessaires à leurs investigations
- Fin des archives circonscrites et datées, ouverture
  - au temps
    - maîtrise avec les technologies d 'aujourd'hui des données nées dans un autre contexte
  - à l 'espace
    - insertion des archives CNES au sein de réseaux ouverts

Le problème de la diversité

- Diversité géographique
  - toutes les données ne sont pas au CNES
  - toutes les données stockées au CNES ne le sont pas au même endroit
- Diversité thématique
  - les données sont toujours structurées selon le point de vue thématique sous lequel on veut les aborder
    - structuration par discipline (ex. : chimie atmosphérique, physique des plasmas)
    - structuration par instrument (ex. : VEGETATION, POLDER)
  - multiplicité des rattachements thématiques
  - diversité des points de vue
- Diversité technologique
  - manifestée par la diversité des formats de données, documents, images

Le cas d 'école du bilan radiatif de la Terre

- Plusieurs instruments et lieux de sockage
  - ERBE et CERES (NASA/Langley)
  - ScaRaB (CNES - LMD)
- Plusieurs formats de données
  - ERBE et ScaRab formats privés
  - CERES format HDF
- Plusieurs méthodes d'accès aux données
  - ERBE et CERES en ligne
  - ScaRab CDROM
- Plusieurs formats documentaires
  - PDF, HTML, PS, .doc
    - selon les documents
    - pour ScaRab, plusieurs formats pour le même document
- Mêmes grandeurs géophysiques mesurées (luminances) **mais**
  - discontinuité des périodes d 'observation
  - faibles recouvrements temporels
- Cas de l'utilisateur qui s'intéresse à une zone géographique déterminée
  - nécessité d 'identifier la zone
    - par ses coordonnées
    - à l 'aide d 'un serveur de nom
  - interrogation des divers lieux de stockage et analyse des résultats sur la base de la documentation
  - extraction et filtrage des données disponibles
  - appel à un serveur de cartes géographiques
  - présentation du résultat du filtrage sur fond de carte

Les technologies disponibles (1/5)

- L 'apport des bibliothèques
  - confrontées au même problème
    - les documents sont aussi complexes que les données
    - l 'échange de documents sous forme numérique est aussi difficile que pour les données
  - solutions proposées
    - outils de balisage des textes qui permettent de localiser et caractériser les éléments structurants d'un texte
      - SGML - DSSSL (XML - XSL)
    - outils de codage des textes par lesquels on rend compte de la sémantique d'un texte muni de son balisage
      - MARC - EAD - TEI
    - des outils d'accès à l'information par lesquels celle-ci est restituée sous forme intelligible à son destinataire
      - Z39.50

Les technologies disponibles (2/5)

- L 'apport du Comité Consultatif des Systèmes de Données Spatiales
  - Langage de description des données EAST
    - description de la structure « logique » des données
      - selon leur type, leur longueur, leur position dans la structure
    - description de la structure « physique des données »
      - selon leur format et le matériel qui les exploite
  - Spécification des dictionnaires de données
    - fondée sur la norme ISO 11179 « *Specification and standardization of data elements* » qui recommande la description d 'une donnée par cinq attributs
      - attributs d'identification (nom, nom abrégé)
      - attributs de définition
      - attributs relationnels (système de classification applicable)
      - attributs de représentation (type de le donnée, formes de représentation valeurs admises)
      - attributs administratifs (tuteur de la donnée)
    - augmentée de règles d 'héritage renforcées permettant de dériver des dictionnaires de produits de données à partir de dictionnaires de communautés

Les technologies disponibles (3/5)

- L ’apport du Comité Consultatif des Systèmes de Données Spatiales
  - Le modèle d ’archivage de données OAIS
    - « *Reference Model for an Open Archival Information System* »
    - composé de six ensembles fonctionnels
      - ingest
      - archival storage
      - data management
      - administration
      - access
      - preservation planning
    - au service de trois entités
      - fournisseur de données
      - utilisateur de données
      - gestionnaire de données
    - fondé sur la notion « d ’information package » où sont distinguées les informations
      - de contenu
      - de représentation
      - de préservation

Les technologies disponibles (4/5)

- Constat : lien étroit entre données et services de données
  - l'accès aux données se fait toujours au travers de services
    - traitements portant sur la structure des données
      - filtrage
      - composition
    - traitements influant sur la signification des données
      - conversion de grandeurs physiques en grandeurs géophysiques
    - analyse des données
      - analyses statistiques
      - présentation des données
      - fouille des données (datamining)
  - Ces services sont tous disponibles au cas par cas
    - sous la forme de développements spécifiques,
    - mais il n'y a pas d'approche globale vraiment opérationnelle
  - Besoin d’approche globale pour les services

Apport des technologies disponibles (5/5)

- L 'apport de la « *Workflow Management Coalition* »
  - Notion de « *Workflow* »
    - défini comme l 'automatisation totale ou partielle d'un processus industriel au cours duquel sont échangés, pour action, des documents, des informations ou des tâches d'un intervenant à l'autre selon un ensemble de règles de procédures
  - Modélisation
    - Workflow Enactment Service (gestion des plans de travaux)
    - Process Definition sur la base d'un « workflow process language definition »
    - Workflow Client Functions (supervision par l 'utilisateur du workflow)
    - Invoked Application Functions (localisation des services)
    - Workflow Interoperability (coopération entre services)
      - chaînage de services
      - inclusion de services dans un autre service
      - services synchronisés par points de rendez-vous
- Apports de L'OGIS et de l'ISO
  - taxinomie des services relatifs à l'Observation de la Terre
  - modélisation des services

Les pièces du puzzle

- Nous disposons à présent :
  - de techniques de codage de l'information documentaire (SGML ou sa variante XML),
  - d'une méthode d'accès à l'information (Z39.50),
  - d'une méthode de description d'enregistrements de données (EAST/OASIS),
  - d'une méthode d'écriture de dictionnaires de données (DEDSL),
  - d'un modèle d'archivage (OAIS),
  - d'un modèle de référence pour la définition des processus (WfMC)

qui nous introduisent au « Bureau des Métadonnées et des Services ».

Le Bureau des Métadonnées et des Services

- Solution proposée en réponse au problème posé
  - Inspirée des « *clearinghouses* » américains et canadien
- Le Bureau des Métadonnées et des Services est
  - un ensemble de serveurs d 'information capable de rendre compte d'un patrimoine de données à l 'intention d'une communauté d 'utilisateurs
- Le Bureau des Métadonnées et des Services offre
  - une clé d'accès
    - somme des connaissances nécessaires à l 'utilisateur pour l'accès à l'information
      - exemple : URL du centre de données gestionnaire
  - un guide
    - fonction de recherche
      - exposé, filtré ou non, de l 'inventaire des données disponibles
    - fonction de présentation
      - pour l 'appréciation, par l 'utilisateur, de l 'adéquation des données à son investigation
  - grâce à des informations de substitution aux « vraies » données
    - ces informations de substitution sont appelées « *métadonnées* »

Les métadonnées - Définitions (1/3)

- Définition synthétique
  - "les données sur les données"

(source ISO)

- Définition fonctionnelle
  - Des données qui permettent
    - d'identifier des données ayant une certaine propriété
    - d'évaluer l'adéquation d'un ensemble de données à un usage déterminé
    - d'acquérir des données préalablement identifiées
    - de traiter et utiliser les données auxquelles on a accès.

(source FGDC)

Les métadonnées - Définitions (2/3)

Définition analytique

- Pour qu'un jeu de données soit "communicable" entre des "parties", il faut que les parties s'accordent sur un "contexte"
- Ce contexte est constitué "d'affirmations a priori" utilisées comme clés d'interprétation des messages échangés entre les parties
  - exemple : le choix des unités de mesure
- Le contexte est habituellement implicite
- Lorsqu'il est explicite, le contexte est rendu par un "schéma conceptuel", dont les éléments sont identifiés à l'aide de données particulières, appelés "métadonnées".

(source F. Bretherton)

Les métadonnées - Définitions (3/3)

Définition philosophique (E .Kant)

- Nos deux sources de connaissances sont
  - l'intuition, capacité à recevoir les représentations d'un objet
  - le concept, qui permet de penser un objet dans son rapport à ces représentations, d'exercer un jugement. Le concept est la connaissance médiate d'un objet.
- Le concept naît d'un travail de l'esprit appelé "synthèse"
  - consistant à lier une diversité d'éléments.
  - Son contraire, ou "analyse", présuppose la synthèse.
    - (on ne peut délier par l'esprit ce qui n'a pas été d'abord lié par lui).
- On appelle "métadonnées" toute représentation numérique d'un concept attaché à un objet, lui-même représenté par des données.
- Comme représentations d'un concept, les métadonnées sont une médiation entre un sujet, exerçant ses facultés cognitives, et l'objet représenté par les données.
- De "bonnes" métadonnées sont en accord avec les structures d'ordre selon lesquelles l'utilisateur exerce son discernement : nécessité du consensus.

Kant (1/2)

- "Notre connaissance jaillit de deux sources principales de notre esprit, la première étant la capacité d'accueillir les représentations (la réceptivité des impressions), la seconde la faculté de connaître un objet par ces représentations (la spontanéité des concepts) ; par la première, un objet nous est donné ; par la seconde, il est pensé en relation à cette représentation (en tant que détermination de l'esprit). Intuition et concepts sont les éléments de toute notre connaissance, de telle sorte que ni les concepts sans une intuition qui leur corresponde d'une certaine façon, ni une intuition sans concepts ne peuvent fournir de connaissance." (Critique de la Raison Pure, 2ème partie : Logique transcendantale, Introduction: Idée d'une Logique transcendantale).

Kant (2/2)

- "Toutes les intuitions, en tant que sensibles, reposent sur des affections, et donc les concepts sur des fonctions. J'entends par fonction l'unité de l'acte par lequel diverses représentations sont rangées sous une représentation commune. Les concepts se fondent donc sur la spontanéité de la pensée, de la même manière que des intuitions sensibles sur la réceptivité des impressions. De ces concepts, l'entendement ne peut faire d'autre usage que de juger par leur moyen. Or comme aucune représentation ne renvoie directement à l'objet, à l'exception de l'intuition, ainsi un concept ne renvoie jamais directement à un objet, mais à quelque autre représentation de celui-ci (que cette représentation soit une intuition ou déjà même un concept). Le jugement est ainsi la connaissance indirecte d'un objet, donc la représentation d'une représentation de cet objet. Dans chaque jugement il y a un concept qui en représente beaucoup d'autres et qui, dans cet ensemble, comprend une représentation donnée, laquelle enfin renvoie directement à l'objet... Penser, c'est connaître par concepts. Et les concepts renvoient, en tant que prédicats de jugements possibles, à quelque représentation d'un objet encore indéterminé."
(Critique de la Raison Pure, 2ème partie: Logique transcendantale, Première Division: Analytique transcendantale, Livre premier: Analytique des concepts)

Spécification de formats de métadonnées

- Schéma de métadonnées
  - traduit la cartographie des concepts applicables à l'objet représenté par le jeu de données
    - via une taxinomie
  - spécifié par BNF, UML, DTD SGML/XML, ...
- Thésaurus
  - instrument de maîtrise du vocabulaire
  - régi par la norme ISO 2788
- Dictionnaire
  - rend compte de la sémantique du schéma de métadonnées
  - régi par les normes ISO 11179 ou CCSDS "DEDSL"
- Règles de définition de "profils"
  - restrictions/extensions d'un format en fonction de besoins spécifiques
- Protocole d'accès "interopérable" à l'information (facultatif)

Un format vraiment universel devient norme ISO

"Interopérabilité"

- Définitions
  - CEOS
    - "Interoperability is the ability of two or more software components to operate reciprocally to overcome distributed resource access barriers imposed by heterogeneous processing environments and heterogeneous data."
    - 4 niveaux d'interopérabilité
  - ISO 19119 (Brodie 1992)
    - "Two components X and Y can operate (are interoperable) if X can send requests R_i for services to Y, based on a mutual undrestanding of R_i by X and Y, and if Y can simillarly return responses S_i to X."
  - OGIS
    - "*Geodata interoperability* refers to the ability to
      - 1) freely exchange all kinds of spatial information about the Earth and about the objects and phenomena on, above and below the Earth'surface; and
      - 2) cooperatively, over networks, run software capable of manipulating such information."
- Exemple de protocole d'interopérabilité : ISO 23950 (Z39.50)

Exemples de normes de métadonnées

- Normes documentaires
  - BIB-1
  - MARC (Machine Readable Catalogue)
  - EAD (Encoded Archival Description)
  - TEI (Text Encoding Initiative)
- Normes généralistes
  - Dublin Core
  - GILS(Government/Global Locator Information Service)
- Normes scientifiques
  - DIF (Directory Interchange Format)
- Normes dédiées à l'OT
  - CSGDM (Content Standard for Digital Geospatial Metadata)
  - CIP (Catalogue Interoperability Protocol)
  - ISO TC211 19115 (Metadata)

La norme DUBLIN CORE

Coverage	caractéristiques spatio-temporelles de la ressource
Description	description textuelle de la ressource (ex : abstract)
Type	catégorie à laquelle appartient la ressource (ex : page Web)
Relation	lien vers une autre ressource, en lien avec la présente
Source	identité d'une autre ressource dont dérive la présente
Subject	thème de la ressource
Title	nom donné à la ressource par son créateur ou son éditeur
Contributor	entité ayant contribué à l'élaboration de la ressource (ex : illustrateur)
Creator	entité,auteur principal du contenu intellectuel de la ressource
Publishor	entité en charge de la publication de la ressource (ex : éditeur)
Rights	définition (ou lien vers la définition) des droits associés à la ressource
Date	date associée à la ressource (ex : date de publication)
Format	format associé à la ressource, en vue de son exploitation
Identifier	chaîne alphanumérique identifiant de la ressource
Language	langue en laquelle le contenu de la ressource est accessible

La norme ISO 19115 - Metadata

- Conçue pour la description de données géographiques
  - Identification — format, browse graphic, usage, constraints, keywords, maintenance
  - Constraint — legal constraints, security constraints
  - Data Quality — lineage, completeness, logical consistency, positional accuracy, thematic accuracy, temporal accuracy)
  - Maintenance — scope and frequency of update
  - Spatial Representation — grid spatial representation, vector spatial representation
  - Reference System — projection parameters, ellipsoid parameters
  - Content Information — coverage description
  - Portrayal catalogue — information on portrayal catalogue used
  - Distribution Information — distributor, medium, transfer options
  - Metadata Extension — user specified extension
  - Application Schema — application schema used to build the dataset
- Très complète - 411 champs disponibles
- Spécification UML
  - dérivation possible d'une DTD XML

Conclusion

- Les métadonnées matérialisent la relation entre
  - un gisement de savoir et
  - un être humain en quête de savoir, doté à cette fin d'une faculté de connaissance qui est l'intellect

    « *en lequel les formes des réalités sensibles ont une existence plus parfaites que dans les réalités sensibles : elles sont alors plus simples et ont plus d'extension* »

    (St Thomas)
- La question est ainsi ramenée à celle de la lecture d'un recueil d'informations et de son interprétation par le lecteur.
  - L'interprétation n'est possible que s'il prévaut, entre le lecteur et le recueil, une certaine connivence.
  - Celle-ci apporte la preuve que le recueil d 'informations est véritablement ordonné à sa finalité.