Les Métadonnées

Une technique au service
de la vision patrimoniale des données

Comment restituer l'intelligibilité d'une information ?

Image de la pièce d'archéologie ...

- La conservation d'une pièce d'archéologie dans un musée ne suffit pas pour permettre son analyse par des spécialistes.
- Il faut conserver aussi des informations telles que :
  - description du lieu de la découverte (nature du terrain, profondeur, ...)
  - description de l'environnement de la pièce (présence d'autres objets, ...)

  Ces informations sont structurées conformément aux règles édictées par la profession.

... transposée aux données numériques

- La conservation en médiathèque de données numériques ne suffit pas pour permettre leur analyse par des spécialistes.
- Il faut conserver aussi des informations telles que :
  - l'organisation des données (formats des fichiers et des enregistrements, ...)
  - la signification des données (nature des paramètres géophysiques restitués, ...)
  - les conditions d'emploi des données (indices de qualité, droits de propriété, ...)

Restituer la connaisance attachée aux données : un enjeu

Constat d'une évolution :

- Transition des systèmes traditionnels ...
  - de traitement des données
  - d'archivage des données
  - de diffusion des données
- ... vers des systèmes au service de la connaissance via des données
  - de plus en plus volumineuses
    - augmentation du nombre des instruments
    - croissance des flux de télémesure
  - de plus en plus complexes
    - complexité des algorithmes
    - multiplication des traitements
  - de plus en plus diversifiées
    - le même ensemble de données peut rendre compte de plusieurs paramètres géophysiques

L'utilisateur potentiel de données ne s'intéresse pas aux données mais à l'information dont elles sont le support (donnée $\neq$ information)

Restituer la connaissance attachée aux données (suite)

Cette évolution résulte de deux facteurs :

- fin des systèmes conçus en fonction des seuls besoins d'une communauté bien identifiée, informée des conditions d'utilisation des données, dont elle a supervisé la définition, au profit de systèmes ouverts au grand nombre
- nécessité de conserver les données sur le très long terme
  - lorsque les données sont la trace d'événements uniques ou dont l'observation ne pourra être renouvelée
  - lorsque les données font partie de séries temporelles longues
  - lorsque des textes réglementaires font obligation de conserver les données

Conséquences sur les archives

- retrait d'une donnée archivée : rarissime
- nécessité pour l'archiviste de composer avec les ruptures technologiques
  - ⇒ développement par le CCSDS d'un modèle d'archivage "ouvert" (modèle OAIS)
- nécessité pour l'archiviste de savoir rendre compte de l'information dont les données sont le support, c–à–d d'en restituer l'intelligibilité

Métadonnées : médiation entre utilisateur et données

Essai de définition

- Restituer l'intelligibilité d'un ensemble de données suppose la maîtrise des schémas de représentation du monde réel, élaborés par un être humain, selon lesquels est structurée l'information portée par les données
- Ces schémas opèrent la médiation entre
  - l'utilisateur à la recherche d'une information
  - les systèmes de données qui hébergent cette information
- A chaque donnée est associé au moins un schéma de représentation
- Le schéma de représentation doit épouser les structures intellectuelles de l'utilisateur :
  - il manifeste les catégories de discernement de l'utilisateur
- Ubiquité du schéma de représentation (pas de colocation avec la donnée)
- Un schéma peut être implicite (oralité, littérature "grise")
- Un schéma peut être explicite (écrit "papier" ou "numérique" formel)
  - sous forme numérique, on l'appelle "métadonnées" (ou "métainformation")

Métadonnées : médiation entre utilisateur et données (suite)

En pratique ...

- Une métadonnée est une description appropriée à son objet, dont elle restitue la syntaxe et la sémantique sous une forme compréhensible par son destinataire
  - syntaxe : règles d'agencement des éléments constitutifs d'un ensemble de données (formats, ...)
  - sémantique : signification d'un ensemble de données (repères thématiques, ...)
- Le même ensemble de données pourra être décrit par plusieurs métadonnées distinctes, selon le profil de ses destinataires
- Détermination par des spécialistes d'un domaine, sur la base du consensus, d'un canevas "type" de métadonnées applicable à tous les ensembles de données du domaine considéré
- Un tel canevas est appelé "format de métadonnées"
- La forme achevée du consensus conduit à l'élaboration d'un format de métadonnées sous les traits d'une norme internationale (ou "standard")

Métadonnées : médiation entre utilisateur et données (suite)

Eléments constitutifs d'un format de métadonnées

- Schéma de représentation
  - issu d'un processus de modélisation d'une partie du monde réel ("Universe of discourse"), par abstraction de concepts
  - donné sous la forme d'une grammaire formelle spécifiant le schéma de concepts
  - à l'aide d'un langage de description de concepts ("Conceptual Schema Language")
- Dictionnaire
  - rend compte du vocabulaire utilisé
  - est formellement défini par la norme ISO 11179 ("Specification and Standardization of Data Elements"), complétée par le CCSDS ("Data Entity Dictionary Specification Language")
- Thesaurus
  - contrôle le vocabulaire applicable
  - est formellement défini par la norme ISO 2788 ("Guidelines for the establishment and development of monolingual thesauri")
- Règles de dérivation de profils (restriction / généralisation du format) au service de communautés spécifiques

Les grandes normes internationales de métadonnées (1/2)

Dublin Core Metadata Initiative (DCMI) – version de base

- "small language for making a particular class of statements about resource"
  "metadata pidgin for digital tourists"
- conçu pour le description synthétique de resources documentaires à l'aide de 15 catégories descriptives réparties en 3 classes :
  - Content
    - Coverage, Description, Type, Relation, Source, Subject, Title
  - Intellectual Property
    - Contributor, Creator, Publisher, Rights
  - Instantiation
    - Date, Format, Identfier, Language

Dublin Core Metadata Initiative (DCMI) – version avec "qualificateurs"

- élément supplémentaire : "audience"
- certains éléments de base peuvent être "qualifiés", par exemple :
  - "Description" peut être qualifié comme "Table_ of_ Contents" ou "Abstract"

Les grandes normes internationales de métadonnées (2/2)

Directory Interchange Format (DIF)

- conçu sous l'autorité du CEOS/WGISS
- maintenu par la NASA dans le cadre du "Global Change Master Directory"
- bien adapté à la description de premier niveau : *~100 catégories descriptives*

Content Standard for Digital Geospatial Metadata (CSDGM)

- conçu et maintenu par le FGDC sous l'autorité du gouvernement américain
- norme officielle aux USA pour la description obligatoire et détaillée d'ensembles de données géographiques : *~340 catégories descriptives*
- utilisé par d'autres pays (Canada)
- intégré par les industriels dans leurs produits (ex. : ESRI)

ISO 19115 – Geographic Information / Geomatics – Metadata

- disponible comme "International Standard" de l'ISO (ISO 19115:2003)
- plus riche que la norme CSDGM : *~410 catégories descriptives*
- élément de la famille de normes géographiques du TC211 de l'ISO

Des métadonnées organisées en Bureau des Métadonnées

Définition d'un Bureau des Métadonnées

- collection de métadonnées
- agencées en base de (méta) données
- offrant une ouverture globale sur des ensembles de données

Rôle

- informer l'utilisateur à propos des données disponibles dans un périmètre déterminé, en lui présentant les métadonnées qui leur sont associées, selon une approche guidée (présentation structurée de l'existant)
- permettre à l'utilisateur de discerner dans la masse des données disponibles celles qui sont véritablement utiles à son investigation, en lui restituant les seules informations nécessaires à l'emploi optimal des données du point de vue de l'utilisateur (appréciation de l'existant)
- contribuer à la maturité d'une thématique (normalisation des concepts)

Le bureau des métadonnées est un serveur d'information, non un serveur de données. **Il est un instrument de promotion et de discernement.**

Les Bureaux de métadonnées existants

Aujourd'hui

- Clearinghouses
  - réseau du "National Spatial Data Infrastructure" (NSDI) aux USA, métadonnées au format CSDGM
  - clearinghouse canadien (GeoConnections), métadonnées au format CSGDM
  - clearinghouse australien, métadonnées au format ANZLIC (version simplifiée du CSDGM)
- Global Change Master Directory
  - réseau de serveurs de métadonnées au format DIF, constituant le "International Directory Network")
  - opérationnel (~12.000 ensembles de données référencés)

Demain

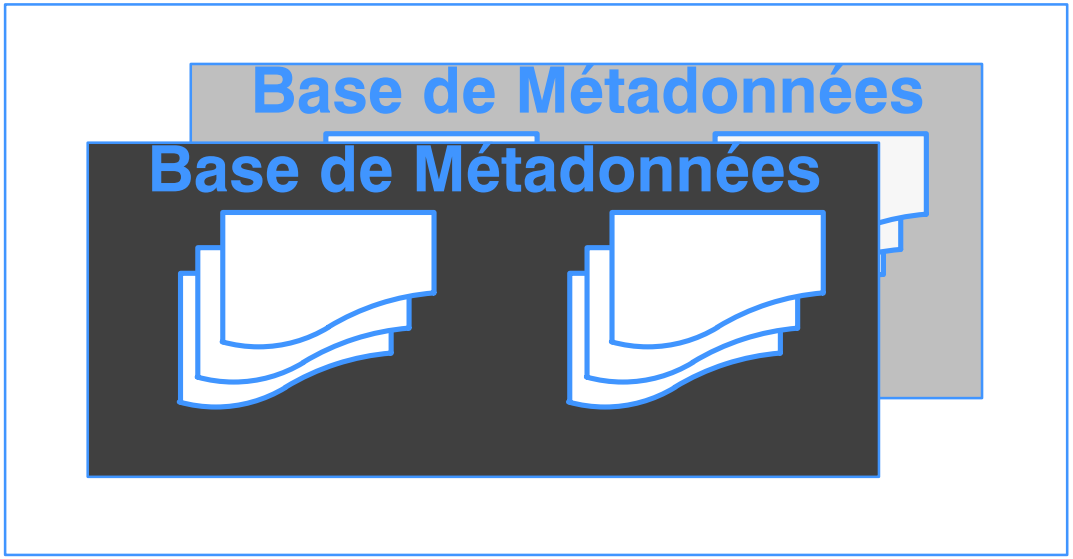
- Extension vers la définition des services applicables aux ensembles de données (cf. les travaux ISO 19119 et OGIS), par exemple :
  - *mapping services*
  - *coverage services*
  - *feature services*

Anatomie d'un Bureau des métadonnées

Administration

Métadonnées

- composition
- validation
- insertion
- retrait
- masquage



Gestion

Définitions

- Profils d'utilisation
- Mots clés
- Thésaurus
- Critères de sélection, p.ex
  - mots clés
  - zone géographique
  - période
  - entité géographique
- Règles de présentation

Consultation

Sélection de métadonnées

- Choix d'un profils d'utilisation
- Choix de critères de sélection
- Choix d'une règle de présentation

Visualisation de métadonnées

Extraction de métadonnées

Technologies utilisées

Interopérabilité :
- Z39.50, SOAP

Encodage de métadonnées :
- SGML, XML, XMLSchema
- éditeurs structurés

Architecture : 3 tiers

Conclusions

- Un Bureau des Métadonnées peut rendre compte avec un grand niveau de détail du patrimoine de données issues de l'observation de la terre, auquel un organisme comme le CNES a contribué.
- A ce titre, il est une réponse à l'action 16 du Plan Stratégique du CNES :
  - "Soutenir la communauté scientifique spatiale nationale durant toutes les phases du projet ..."
    - "... le CST facilitera l'accès des équipes scientifiques aux données spatiales."
- Son efficacité sera réelle s'il sait s'intégrer, par le respect de normes reconnues, dans un réseau international de "clearinghouses".
- Sa pérennité sera assurée s'il y a une réelle volonté institutionnelle en ce sens. Le succès de la norme américaine CSDGM s'explique ainsi :

Executive Order 12906, "Coordinating Geographic Data Acquisition and Access : The National Satial Data Infrastructure"
"Beginning nine months from the date of this order, each agency shall document all new geospatial data it collects and produces, either directly or indirectly, using the standard under development by the FGDC, and make that standardized documentation electronically accessible to the Clearinghouse network. Within one year of the date of this order, agencies shall adopt a schedule, developed in conjunction with the FGDC, for documenting, to the extent practical, geospatial data previously collected or produced, either directly or indirectly, and making that data documentation electronically accessible to the Clearinghouse network." (signé en 1994 par le Président Clinton)

Références

- Dublin Core : http://dublincore.org
- GCMD/DIF : http://gcmd.nasa.gov
- FGDC/CSDGM : http://fgdc.gov/metadata/meta_stand.html
- ISO : http://www.iso.ch
- Open GIS : http://www.opengis.org