CENTRE NATIONAL D'ETUDES SPATIALES

La question des Métadonnées au CNES

Origine de la question

Confrontation du CNES à la nécessité de préserver ses données sur le long terme

- les données sont souvent “uniques” car difficiles à reproduire ou témoins d’un phénomène observé qui ne se reproduira pas
- les données font souvent partie de séries temporelles nécessaires pour des études à long terme (climatologie, ...)
- les données anciennes peuvent être utiles, éventuellement associées à d’autres données, à des fins différentes de celles pour lesquelles elles ont été conçues

Évolution des segments sol de traitement, archivage, diffusion de données

- fin des systèmes seulement circonscrits aux besoins de communautés bien définies, aux services desquelles ils ont été construits
- mise en place de systèmes au service de la connaissance
  - gestion du temps : prise en compte du long terme, réutilisation avec la technologie d’aujourd’hui de données crées dans un autre contexte
  - gestion de l’espace : permettre l’utilisation conjointe de données dispersées en divers lieux

L'analyse du CNES

Importance du contexte

- L'utilisation de données est facile lorsque sont disponibles
  - leur contexte technique (forme)
  - leur contexte thématique (sens)
- Lorsque l'un ou l'autre des contextes est perdu ou s'il n'est plus possible d'en rendre compte, l'utilisation fiable des données est impossible !

La menace du temps

- Contexte technique et contexte thématique sont toujours disponibles au début
  - de manière formelle
    - dans des documents "papier" ou "numériques"
    - par l'intermédiaire de dispositifs techniques appropriés
  - de manière informelle, dans la mémoire des individus
- Au fil du temps, la disponibilité du contexte technique et/ou du contexte thématique s'atténue si ne sont pas prises des mesures en faveur de sa préservation : éviter l'oubli, composer avec les ruptures technologiques.

Les actions entreprises par le CNES

Préserver les données "physiques"

- opérationnel au CNES
- pris en charge au Centre Informatique par le Service de Transfert et d'Archivage des Fichiers (STAF)

Préserver le contexte technique

- opérationnel au CNES
- pris en charge par la technologie EAST/OASIS de description des données
  - développée par le CNES, endossée par le CCSDS et l'ISO (norme ISO15889:2003)
  - dotée de logiciels d'implantation

Préserver le contexte thématique

- non opérationnel au CNES
- expérimenté par des travaux technologiques
- approché dans le cadre de la normalisation CCSDS des dictionnaires de données

Préserver le contexte thématique

Enjeu

Un organisme comme le CNES doit être capable de rendre compte du patrimoine de jeux de données issus de la télémesure des instruments qu'il a développés directement ou dans le cadre de coopérations.

Pour être recevable par le plus grand nombre, un jeu de données doit être :

- identifié (titre, mots-clés, couverture spatio-temporelle,...)
  - sinon on ne peut pas savoir qu'il existe
- convenablement décrit (caractéristiques, objet, qualité, référentiel d'emploi,...)
  - s'il est mal ou insuffisamment décrit aux yeux de l'utilisateur potentiel, il sera ignoré ou rejeté par celui-ci
- utilisable
  - il faut savoir où le trouver
  - il faut savoir y accéder
  - il faut savoir en tirer parti

Rendre compte du contexte thématique

De la donnée à l'information

- il n'y a pas grande différence entre un jeu de données et un document
  - "Un document est une preuve à l'appui d'un fait" (Suzanne Briet, 1951)
  - Les données sont une "représentation formelle de faits, concepts ou instructions convenant à la communication, l'interprétation ou le traitement par des êtres humains ou des automatismes". (ISO11179 "Specification and standardization of data elements")
- Jeux de données et documents sont porteurs d'information
  - information : "connaissance relative à des objets tels que des faits, événements, choses, processus ou idées, y compris des concepts, laquelle, dans un certain contexte, a une signification particulière" (ISO11179)
  - objet : "élément de la réalité qui peut être conçu ou perçu" (ISO1087 "Terminology – Vocabulary)
  - concept : "unité de pensée constituée par abstraction sur la base des caractéristiques communes à un ensemble d'objets" (ISO1087). Un concept peut être superordonné ou subordonné à d'autres concepts dans une relation dans une relation générique ou partitive, ou coordonné avec d'autres concepts.

Rendre compte du contexte thématique

De l'information au jugement

- *Notre connaissance émane de deux sources principales de notre âme, dont la première est la capacité d'accueillir les représentations (la réceptivité des impressions), la seconde la faculté de reconnaître un objet par ces représentations (la spontanéité des concepts) ; par la première, un objet nous est donné, par la seconde, il est pensé en rapport avec cette représentation (en tant que pure détermination de l'âme). Intuition et concepts constituent ainsi 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 concept ne peuvent fournir de connaissance.*
  (Emmanuel Kant, Critique de la Raison Pure, Théorie transcendantale élémentaire, 2ème partie : La logique transcendantale, Introduction : Idée d'une logique transcendantale, I. : De la logique en général)
- *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 ordonnées sous une représentation commune. Les concepts se fondent donc sur la spontanéité de la pensée, comme les intuitions sensibles sur la réceptivité des impressions. De ces concepts, l'entendement ne peut alors faire d'autre usage que de juger par leur moyen. Or comme aucune représentation ne va imédiatement à l'objet, à l'exception de l'intuition, ainsi un concept n'est jamais rapporté immédiatement à un objet, mais à quelque autre représentation de celui-ci (qu'elle soit intuition ou qu'elle soit elle-même déjà concept). Le jugement est ainsi la connaissance médiate d'un objet, par conséquent la représentation d'une représentation de celui-ci. Dans chaque jugement il y a un concept qui compte pour beaucoup d'autresconcepts et qui, parmi ceux-là, comprend une représentation donnée, laquelle est rapportée immédiatement à l'objet... Penser, c'est connaître par concepts. Et les concepts sont rapportés, en tant que prédicats de jugements possibles, à quelque représentation d'un objet encore indéterminé.* (Emmanuel Kant, Critique de la Raison Pure, Théorie transcendantale élémentaire, 2ème partie : La logique transcendantale, 1ère division: L'analytique transcendantale, 1er livre : L'analytique des concepts)

Rendre compte du contexte thématique

La modélisation conceptuelle

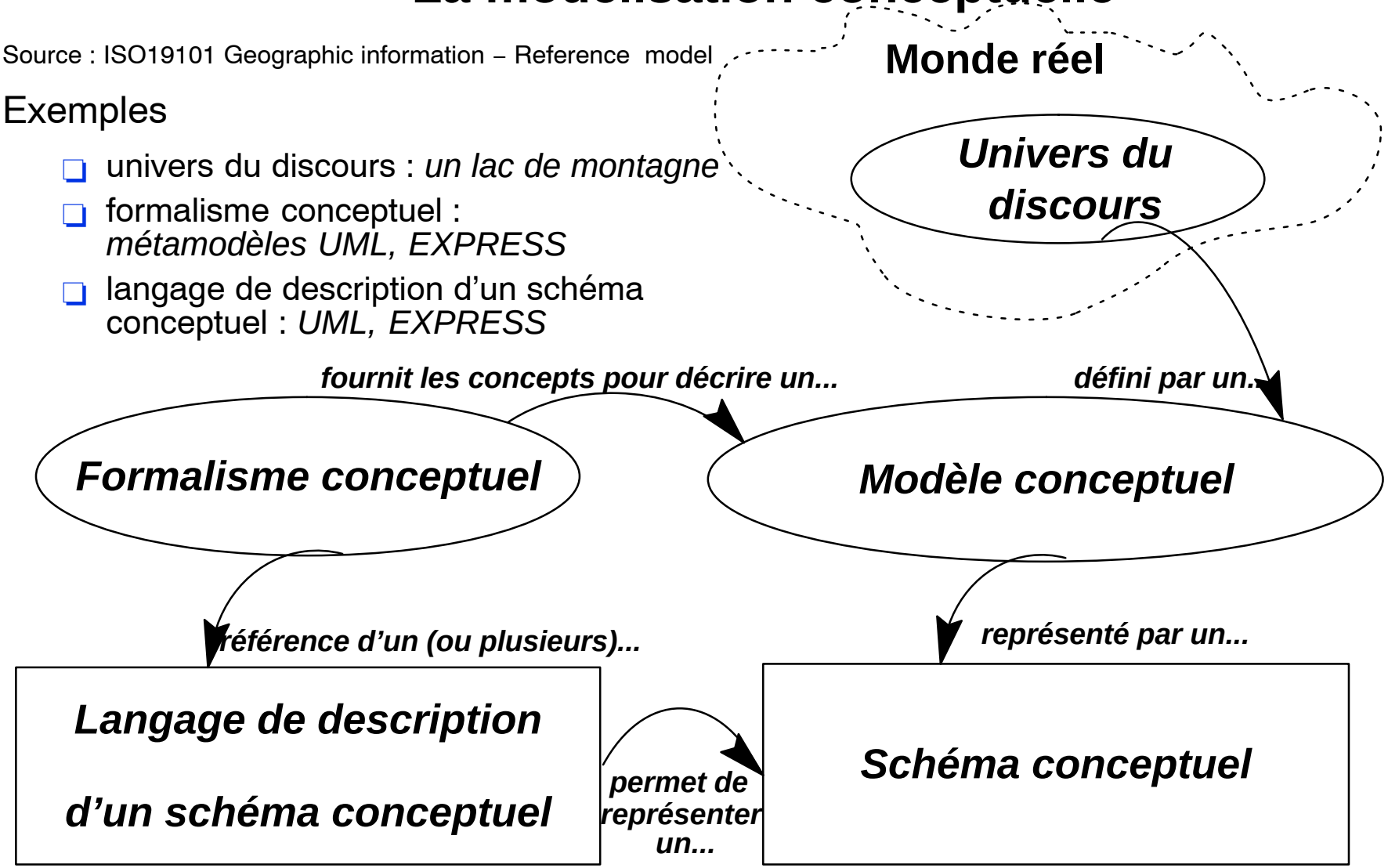
- Rendre compte du contexte thématique d'un jeu de données, c'est rendre explicites les concepts qui en constituent l'armature.
- Tout jeu de données ou document manifeste un "certain regard" sur le monde réel ; ce regard dépend du modèle conceptuel appliqué au monde par l'observateur.

La modélisation conceptuelle

Source : ISO19101 Geographic information – Reference model

Exemples

- univers du discours : *un lac de montagne*
- formalisme conceptuel : *métamodèles UML, EXPRESS*
- langage de description d'un schéma conceptuel : *UML, EXPRESS*



La modélisation conceptuelle

L'univers du discours est un fragment du monde réel ou d'un monde hypothétique que l'on souhaite modéliser. Le résultat de la modélisation est un modèle conceptuel.

L'univers du discours ne contient pas seulement des entités "physiques". Il contient aussi, éventuellement, leurs propriétés, leurs fonctions, les relations qui peuvent exister entre plusieurs entités.

Le modèle conceptuel est construit à l'aide des concepts proposés par un formalisme conceptuel (il peut y avoir plusieurs formalismes conceptuels pour le même univers du discours).

Le modèle conceptuel peut rester intellectuel. Il peut aussi être représenté par un schéma conceptuel exprimé à l'aide d'un langage de description spécifique, capable de rendre compte du formalisme conceptuel.

"C'est cette représentation d'un procédé général de l'imagination, servant à procurer à un concept son image, que j'appelle le schéma de ce concept." (Critique de la Raison Pure, 1ère partie : Analytique des concepts, Livre deuxième : Analytique des principes, Chapitre 1 : Du schématisme des concepts purs de l'entendement).

La modélisation conceptuelle

Application à l'observation de la terre

Lorsque l'univers du discours est la terre (et son environnement), nous pouvons utiliser un formalisme conceptuel comme celui-ci :

- une entité est désignée par son nom (ex : maison) et caractérisée par
  - sa définition (ex : bâtiment destiné à l'habitation d'une famille)
  - ses propriétés
  - ses contraintes (ex : hauteur maximale)
  - ses relations avec d'autres entités (ex : cas particulier d'habitation)
- une propriété peut être
  - une opération applicable (ex : élévation)
  - un attribut propre (ex : hauteur)
  - une association (ex : mitoyenneté avec un immeuble)
- un attribut est caractérisé par son type
  - thématique (ex : style de la maison)
  - temporel (ex : date de construction)
  - localisation géographique (ex : nom de la ville)

La modélisation conceptuelle

Ce formalisme permet de modéliser l'entité géographique constituée (comme suggéré plus haut) d'une maison.

Mais le même formalisme peut s'appliquer à quelque chose de plus immatériel comme l'entité géographique constituée par le vent, au-dessus d'une zone particulière de l'océan.

En pratique, nous avons accès à l'univers du discours par l'intermédiaire de perceptions, observations, mesures (cf. la citation d'E. Kant sur l'intuition) filtrées conformément à un modèle conceptuel. L'enregistrement de celles-ci n'abolit pas ce modèle conceptuel mais l'enregistrement est lui-même tributaire d'un modèle conceptuel qui peut, à son tour, être rendu par un schéma conceptuel. Ce schéma porte le nom de schéma d'application (ISO 19101).

Un schéma d'application est donc le schéma d'un jeu de données. Il rend compte de façon complète et précise du contenu et de la structure d'un jeu de données.

La modélisation conceptuelle

Modèles et schémas conceptuels sont des appareils qui nous permettent d'exprimer des jugements ("les concepts sont des prédicats de jugements possibles").

Un jugement est valable s'il est recevable par d'autres (sinon, il est seulement l'expression d'une opinion). Pour cela, modèles et schémas conceptuels à partir desquels le jugement est rendu doivent être ceux d'une communauté. Cette adhésion s'exprime par des normes (de droit ou de fait).

La normalisation est l'activité de synthèse par lequel une communauté exprime **consciemment** les concepts qui la fondent comme communauté :

*Seule, la liaison (*conjuctio*) d'une diversité d'éléments en général ne peut jamais venir en nous par l'intermédiaire des sens et ne peut donc pas non plus être contenue en même temps dans la forme pure de l'intuition sensible car elle est un acte de la spontanéité de la faculté de représentation; et, comme il faut l'appeler entendement, pour la distinguer de la sensibilité, ainsi toute liaison, que nous en ayons conscience ou non, qu'elle soit liaison d'une diversité d'éléments de l'intuition ou de concepts et, s'agissant de l'intuition, que celle-ci soit sensible ou non, est un acte de l'entendement que nous désignerons du nom commun de synthèse, afin de faire remarquer ainsi que nous ne pouvons alors rien nous représenter comme lié dans l'objet si nous ne l'avons pas lié nous-même auparavant, et que, de toutes les représentations, la liaison est la seule qui ne puisse nous être fournie par des objets, mais seulement rapportée par le sujet lui-même, parce qu'elle est un acte de sa spontanéité. On remarquera sans peine ici que cet acte doit être originellement unique et équivalent pour toute liaison, et que la décomposition, l'analyse, qui semble être son contraire, la présuppose toujours ; car où l'entendement n'a rien lié auparavant, il ne saurait non plus rien décomposer, puisque c'est par lui seul que ce qui a été lié a pu être donné comme lié à la faculté de représentation.* (Emmanuel Kant, Critique de la Raison Pure, Théorie transcendantale élémentaire, 2ème partie: La logique transcendantale, 1ère division : L'analytique transcendantale, 1er livre : L'analytique des concepts, 2ème section : De la déduction des concepts purs de l'entendement, 2ème paragraphe : De la déduction transcendantale des concepts purs de l'entendement : De la possibilité en général d'une liaison).

Les métadonnées

Les métadonnées (*metadata*) sont des données particulières. Elles sont habituellement agencées en jeux de métadonnées (metadata datasets).

Un jeu de métadonnées rend compte d'un ou plusieurs jeux de données, en tant que jeu de données. La manière dont un jeu de métadonnées rend compte d'un jeu de données en tant que jeu de données est déterminée par le modèle conceptuel (et le schéma conceptuel) dont le jeu de métadonnées est tributaire.

Ce modèle conceptuel est l'appareil qui permet d'exprimer un jugement sur un jeu de données en tant que jeu de données. Un tel jugement permet éventuellement de distinguer un jeu de données parmi d'autres jeux de données. Un jeu de métadonnées est ordinairement un instrument de discernement.

Comme tout jeu de données, un jeu de métadonnées est régi par un schéma d'application.

La normalisation des jeux de métadonnées

Pour que le jugement rendu par un jeu de métadonnées soit valable, c'est-à-dire recevable par d'autres, il faut que son modèle conceptuel (et donc son schéma d'application) soit celui d'une communauté. Cette adhésion s'exprime, comme pour les autres schémas d'application par des normes de droit ou de fait.

Quelques normes de métadonnées (par complexité croissante)

- Dublin Core — Dublin Core Metadata Initiative (DCMI)
- BIB-1 — norme bibliographique internationale (Library of Congress)
- GILS — Government Information Locator Service (USA)
- DIF — Directory Interchange Format (CEOS/GCMD)
- CSDGM — Content Standard for Digital Geographic Metadata (FGDC)
- ISO 19115 — Geographic Information Metadata (ISO)

La norme ISO 19115

- publiée par l'ISO en 2003
- dévolue à la description de l'information géographique
- plus de 400 éléments descriptifs pour un jeu de données, ainsi classés :
  - identification
    - références et description synthétique du jeu de données
    - couverture spatio–temporelle
  - contenu
    - entités géographiques
    - attributs
  - distribution
  - qualité
    - indicateurs de qualité
    - historique du jeu de données
  - représentation spatiale
  - systèmes de référence
  - présentation
  - schéma d'application

Le langage de description de schéma conceptuel est composé de :

- UML (partie statique)
- IDL (types de base)
- OCL
- dictionnaire ISO 11179

Le Bureau des Métadonnées du CNES

Point de vue technique

Le Bureau des Métadonnées est un système d'information composé

- d'une ou plusieurs bases de jeux de métadonnées
  - réparties en plusieurs lieux
  - chaque base est alimentée au niveau local
  - toutes les bases peuvent être interrogées en même temps
- d'interfaces d'accès pour :
  - l'interrogation par l'utilisateur
  - l'administration des jeux de métadonnées
    - ingestion de métadonnées
    - retrait de métadonnées
  - le management global
    - détermination des normes de métadonnées acceptées
    - détermination des profils d'utilisation

Le Bureau des Métadonnées du CNES

Point de vue technique (suite)

La conception du Bureau des Métadonnées suit les recommandations de l'OAIS (Reference Model for an Open Archival Information System du CCSDS)

Ce modèle définit un système d'archivage comme étant la composition d'un ensemble d'unités fonctionnelles pouvant avoir chacune son autonomie :

- ingest — *ingestion des métadonnées*
- archival storage — *stockage des métadonnées*
- data management — *gestion des métadonnées en base*
- administration — *contrôle au quotidien du Bureau des Métadonnées (ex :acceptation des métadonnées)*
- preservation planning — *observation de l'environnement du Bureau des Métadonnées, audits*
- access — *accès aux métadonnées par les utilisateurs*

Les métadonnées sont confectionnées à l'extérieur du système à l'aide d'un éditeur structuré. Les fichiers résultant de l'édition structurée sont ingérés dans le système.

Le Bureau des Métadonnées du CNES

Point de vue thématique

Le Bureau des Métadonnées sert les besoins de l'observation de la terre. L'analyse du besoin a été conduite par la méthode des interviews : des producteurs potentiels de métadonnées comme des utilisateurs potentiels du Bureau des Métadonnées ont été interrogés.

Conclusions de l'analyse du besoin

- approche du Bureau des Métadonnées par profil d'utilisation
  - un profil d'utilisation rend compte d'un domaine (métier, thématique,...) ;
  - un profil d'utilisation est déterminé par un ensemble de mots clés représentatifs du domaine ; ces mots clés sont contrôlés par des thésaurus ;
- faculté d'interroger le Bureau des Métadonnées par mots clés thématiques
- faculté d'interroger le Bureau des Métadonnées par sélection géographique
  - noms géographiques
  - rectangle
- faculté d'interroger le Bureau des Métadonnées par sélection temporelle
  - dates de début et de fin
  - saison
- faculté d'accès à des représentations graphiques (*browse*, *quicklooks*,...)

Le Bureau des Métadonnées du CNES

Point de vue thématique (suite)

A chaque profil sont associés également :

- les critères de sélection pertinents pour le profil
- les règles de présentation à l'utilisateur
  - composition et présentation des *shortlists* résultant d'une interrogation de l'utilisateur
  - présentation des métadonnées à l'utilisateur

Le nombre de profils n'est pas limité. Il est même préférable d'en créer beaucoup pour réduire le bruit (ratio des métadonnées indésirables / total des métadonnées restituées).

Il est possible de créer des profils génériques englobant tous les autres profils.

Le Bureau des Métadonnées gère à part les informations nominatives (ex : adresse d'un organisme). Une modification dans une information nominative est répercutée dans toute présentation de métadonnées.

Le Bureau des Métadonnées donne accès en ligne à des documents adventices (thesaurus, etc.)

Le Bureau des Métadonnées du CNES

Point de vue thématique (suite)

Le Bureau des Métadonnées ne reconnaît qu'une norme à la fois : toutes les métadonnées devant être ingérées par le Bureau des Métadonnées doivent être conformes à cette norme.

Toutefois, le Bureau des Métadonnées peut représenter une partie seulement des éléments de métadonnées restituées à l'utilisateur, par application d'un filtre. En particulier, l'application du filtre peut convertir, sur demande de l'utilisateur, la norme de référence du Bureau des Métadonnées en une norme plus simple.

A ce jour, la norme de référence du Bureau des Métadonnées est la norme

ISO 19115:2003 Geographic data – Metadata

Elle a été choisie car elle est la plus riche de toutes les normes connues au CNES, dans le domaine de l'observation de la terre.

Le Bureau des Métadonnées peut gérer simultanément plusieurs versions de la norme.

Le Bureau des Métadonnées du CNES

Point de vue thématique (suite)

Nous assimilons un jeu de métadonnées à un *document*, c'est-à-dire au *support d'un enseignement* (cf. l'étymologie du mot *document*).

Un jeu de métadonnées doit être pensé en fonction de ses destinataires : les utilisateurs du Bureau des Métadonnées. Il devrait se couler dans le moule des catégories intellectuelles selon lesquelles les utilisateurs du Bureau des Métadonnées conçoivent leur approche du réel.

Il faut se souvenir que l'utilisateur du Bureau des Métadonnées est en situation d'ignorance. Cette situation d'ignorance est le plus souvent une situation pénible. Le caractère pénible de l'ignorance est le principal moteur d'une recherche d'information (Michael Buckland).

Méthode :

La composition d'un jeu de métadonnées devrait suivre les préceptes de l'analyse documentaire : observer, comprendre, structurer, formuler, résumer, indexer
(cf. *L'analyse documentaire*, Suzanne Waller, ADBS Editions, 1999).

Le Bureau des Métadonnées du CNES

Point de vue thématique (suite)

- **observer** : survoler le jeu de données, le situer dans son environnement (autres jeux de données, lieux de stockage, gestionnaires, etc.), identifier sa bibliographie (auteurs, date, langues, citations, etc.), en bref tous les éléments factuels ;
- **comprendre** : discerner la problématique du jeu de données et les catégories d'intérêt pour lesquelles il peut être une réponse ;
- **structurer** : discerner la logique qui a conduit le jeu de données à ce qu'il est, avec ses qualités et ses limites ;
- **formuler** : identifier les éléments de métadonnées qu'il conviendra d'enrichir ; identifier le vocabulaire adéquat ;
- **résumer** : "Le résumé est un texte concis reflétant fidèlement, sans interprétation ni critique, le contenu d'un document. Il a pour but d'aider le lecteur à cerner la pertinence du document, vis-à-vis de l'information recherchée" (AFNOR Z 44-004) ;
- **indexer** : trouver les mots qui inspireront la requête de l'utilisateur, choisis dans un document de contrôle utilisant un langage documentaire : thésaurus (ISO2788), catalogue de noms géographiques (ISO19112), catalogue d'entités géographiques (ISO19110).

Le Bureau des Métadonnées du CNES

Point de vue thématique (suite)

L'expérimentation thématique du Bureau des Métadonnées par le GIP MEDIAS France a mis en lumière la difficulté à composer des jeux de métadonnées porteurs de sens pour les utilisateurs :

- que renseigner dans les métadonnées ?
- quel titre choisir ?
- comment rédiger un résumé ?
- quels sont les mots clés pertinents ? dans quels thésaurus ?
- comment rendre compte de la qualité d'un jeu de données ?

Remarque de l'un des *cobayes* : *la restitution de l'information devrait être en "langage humain"* (cf. le rapport bilan de l'expérimentation).

A notre sens, ces difficultés ont pour origine l'insuffisante modélisation conceptuelle au sein des communautés.

Le Bureau des Métadonnées du CNES

Point de vue thématique (suite)

Le Bureau des Métadonnées est un service d'information. Pour être reconnu comme tel, il devra surmonter six obstacles :

- **identification** : une source d'information doit être identifiable comme réponse possible à une série de questions ;
- **disponibilité** : une source d'information doit être communicable sous une forme appropriée ;
- **prix payé par l'utilisateur** : le prix (en argent, temps, effort, inconfort) doit être acceptable pour l'utilisateur ;
- **prix payé par le service d'information** : le prix (en argent, temps, effort, aménagement) doit être acceptable pour la tutelle du service ;
- **compréhension** : la source d'information communiquée à l'utilisateur doit être comprise de ce dernier ; sinon, il faut lui fournir l'expertise nécessaire ;
- **acceptabilité** : une autorité reconnue doit endosser la source d'information communiquée à l'utilisateur pour que celle-ci soit acceptée.

(cf. *Information and Information Systems*, Michael Buckland, Greenwood Press, 1991)

Retour sur la question des métadonnées au CNES

Point de vue technique

- la technologie est globalement maîtrisée par le CNES

Point de vue thématique

- une phase d'expérimentation thématique du Bureau des Métadonnées par MEDIAS France a conduit à des recommandations d'amélioration du Bureau des Métadonnées, prises en compte dans la version actuelle et dans la version à venir
- une phase de composition opérationnelle de jeux de métadonnées pour le Bureau des Métadonnées est en cours chez MEDIAS France

Point de vue institutionnel

- plusieurs rapports ont été produits en interne et en coopération avec d'autres organismes sur l'intérêt, pour l'établissement, d'une politique "métadonnées"
- pour l'instant, la greffe n'a pas pris
  - mais pas d'obstacle de principe à la poursuite de l'activité
- sans doute parce qu'il s'agit d'une activité "transverse", plutôt en décalage par rapport à la culture "projet" du CNES

Un point de vue de l'étranger

Executive Order 12906
"Coordinating Geographic Data Acquisition and Access : The National Spatial 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)

Questions ?