Gilbert Maudire
(Ifremer)

"Partage des bases de données"

26 Novembre 2013, Ifremer, Brest

Ifremer

Lab-STICC

http://wwz.ifremer.fr/bigdata

Partage et interaction des bases de données

L'exemple de SeaDataNet

sur les bases techniques de la directive Inspire

Enjeux et objectifs

- **Un constat**
  - L'observation de l'océan est très répartie
    en particulier dans les domaines littoral et côtier
  - Environ 1100 organismes actifs pour les mers bordant l'Europe
  - Modes de fonctionnement divers (financement, politique de données, procédures)
- **Faciliter la vie de l'utilisateur des données**
  - Point d'accès unifié à de grands ensembles de données répartis
  - Résultats des requêtes présentés sous forme harmonisée
  - Règles d'accès précisées
- **Garder le contrôle au producteur des données**
  - Visibilité du producteur et de ses données
    maintien de l'effort d'observation, valorisation intellectuelle (voire financière)
  - Améliorer la qualité globale du jeu de données
    - Proximité observation / gestion de données
    - Intérêt à faire valoir ses propres données
    - Connaissance des phénomènes étudiés (zones géographiques, expérience thématique, ...)
- **Renforcement de la visibilité globale de l'observation et possibilité de synthèses (produits)**
  - Vision globale de l'effort d'observation
  - Collections de données de référence
  - Analyses géostatistiques, y compris analyse des manques

➔ **Interopérabilité et système de gestion de données distribué peuvent ils apporter une réponse?**

SeaDataNet
un exemple de système distribué à large échelle

- **Projet d'infrastructure de Recherche (I3) financé par la DG-Recherche**
  - depuis 2006 (2 projets successifs)
  - orienté physique, chimie et biologie de la colonne d'eau
  - environ 50 partenaires, 35 pays
  - plusieurs organisations internationales :
    COI-IODE, JCOMM-OPS, CIEM, EuroGoos, ...

- **Des frères jumeaux**
  - Geo-Seas : Géophysique et Géologie
  - EMODNET (European Marine Observation and Data Network)
    à vocation plus institutionnelle (DG-MARE) :
    Hydrographie, Chimie, ...

Au total

80 centres de données actuellement connectés
dont en France CNRS/INSU/IPG Strasbourg, BRGM, SHOM, Ifremer

L'interopérabilité dans SeaDataNet (1/2)

•Parler le même langage

- Définir des vocabulaires communs (référentiels)
  - Gouvernance : mise à jour, nouvelles listes, correspondances ... vers des ontologies
  - Serveurs d'ontologies (SKOS : Simple Knowledge Organization System) : NERC/BODC
  - Serveurs de taxinomie (ERMS / WORMS)
  - Avec une gouvernance partagée (forums, messagerie) : experts mondiaux
- Adopter des standards
  - Métadonnées (ISO 19115/19139: conformité Inspire, Darwin Core pour la biodiversité)
  - Formats de données (NETCDF, fichiers texte à colonnes, SEG, ...)

•Une harmonisation des données et de la gouvernance

- Qualité des données connue
  - Règle de contrôle commune, échelle de flags de qualité, tests obligatoires
  - En partenariat avec la communauté internationale (programme, COI, WMO)
- Accès à la donnée
  - Les droits peuvent être différents mais sous des règles mises en communs (pour pouvoir les enregistrer dans les métadonnées)
  - Typologie / Droits des utilisateurs (Education, Recherche, Décision publique, Commercial...)
  - Niveau d'accessibilité des données
  - Mécanisme de décision
  - Définition des droits et devoir des producteurs (politique de données)
  - Définition des droits et devoir des utilisateurs (licence d'utilisation des données et produits)

L'interopérabilité dans SeaDataNet(2/2)

•Des services en ligne

- Interopérabilité technique (les services préconisés par Inspire)
  Découverte, Accès aux données, Visualisation des données
  - Catalogues de métadonnées et protocole d'interrogation (OGC CS-W)
  - Protocoles d'échange automatisés
    (ftp, OGC WFS, WCS, famille Sensor Web Enablement, OBIS pour la biologie)
- Gestion des utilisateurs et du support aux utilisateurs (helpdesk)
- Gestion de la disponibilité
  - Monitoring des services
  - Qualité de service et engagement de service

•Des services hors-ligne (arrière boutique)

- Gestion des données au quotidien (formatage, ...), de leur qualité et des outils associés
- Elaboration des métadonnées et outils associés
- Outils de gestion du réseau : utilisateurs, disponibilité, asynchronisme des requêtes longues,...
- Maintenance des systèmes internes
  Suivi des incidents et de leur remédiation
- Statistiques et indicateurs : utilisation, jeux de données accessibles, ...

Comment ça marche?

Portail
(Point d'entrée utilisateur)

Catalogue harmonisé des données

(Métadonnées « ISO 19115 »)

Download Manager → **Base de données 1**

Download Manager ↔ **Base de données 2**

Download Manager → **Base de données i**

Download Manager → **Base de données n**

Comment ça marche?

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Catalogue harmonisé des données

(Métadonnées « ISO 19115 »)

Download Manager → **Base de données 1**

Download Manager ↔ **Base de données 2**

Download Manager → **Base de données i**

Download Manager → **Base de données n**

Comment ça marche?

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Identification des données correspondantes et des bases.

Catalogue harmonisé des données

(Métadonnées « ISO 19115 »)

Download Manager → **Base de données 1**

Download Manager → **Base de données 2**

Download Manager → **Base de données i**

Download Manager → **Base de données n**

Comment ça marche?

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Identification des données correspondantes et des bases.

Catalogue harmonisé des données
(Métadonnées « ISO 19115 »)

Demandes de mise à disposition

Download Manager → **Base de données 1**
Download Manager → **Base de données 2**
Download Manager → **Base de données i**
Download Manager → **Base de données n**

Comment ça marche?

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Catalogue harmonisé des données
(Métadonnées « ISO 19115 »)

Identification des données correspondantes et des bases.

Demandes de mise à disposition

Accord nécessaire si donnée pas en accès libre

Download Manager — **Base de données 1**

Download Manager — **Base de données 2**

Download Manager — **Base de données i**

Download Manager — **Base de données n**

Comment ça marche?

Registre des utilisateurs

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Identification des données correspondantes et des bases.

Catalogue harmonisé des données

(Métadonnées « ISO 19115 »)

Demandes de mise à disposition

Accord nécessaire si donnée pas en accès libre

Download Manager — Base de données 1

Download Manager — Base de données 2

Download Manager — Base de données i

Download Manager — Base de données n

Comment ça marche?

Registre des utilisateurs

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Catalogue harmonisé des données
(Métadonnées « ISO 19115 »)

Identification des données correspondantes et des bases.

Demandes de mise à disposition

Renvoi de la donnée après accord et harmonisation

Accord nécessaire si donnée pas en accès libre

Download Manager

Base de données 1

Download Manager

Base de données 2

Download Manager

Base de données i

Download Manager

Base de données n

Comment ça marche?

Registre des utilisateurs

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Identification des données correspondantes et des bases.

Catalogue harmonisé des données

(Métadonnées « ISO 19115 »)

Demandes de mise à disposition

Renvoi de la donnée après accord et harmonisation

Accord nécessaire si donnée pas en accès libre

Génération des métadonnées :

Outils Mikado, GeoNetwork, ...

Download Manager

Base de données 1

Download Manager

Base de données 2

Download Manager

Base de données i

Download Manager

Base de données n

Comment ça marche?

Registre des utilisateurs

Requête
Quoi, Où, Quand, Qui, ...

Portail
(Point d'entrée utilisateur)

Identification des données correspondantes et des bases.

Catalogue harmonisé des données
(Métadonnées « ISO 19115 »)

Demandes de mise à disposition

Renvoi de la donnée après accord et harmonisation

Accord nécessaire si donnée pas en accès libre

Génération des métadonnées :
Outils Mikado, GeoNetwork, ...

Download Manager

Download Manager

Download Manager

Download Manager

Base de données 1

Base de données 2

Base de données i

Base de données n

Serveur de vocabulaires commun

Les métadonnées

- **Un ensemble de catalogues communs de l'observation aux produits**

Common Vocabularies

Observations

Cruise Summary Reports
CSR

Permanent Observing Systems
EDIOS

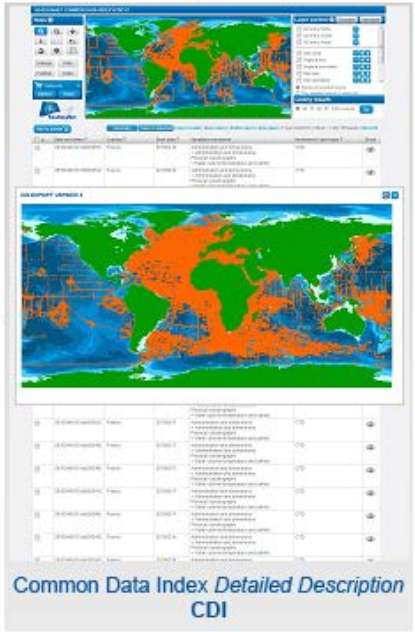
Marine Data Sets *General Descriptions*
EDMED

Data management

Marine Research Organizations
EDMO

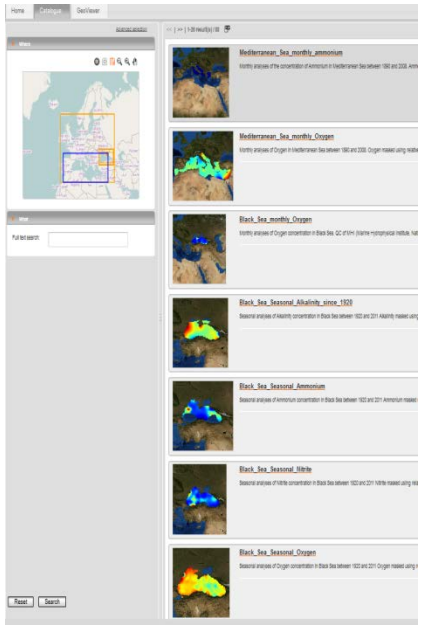
Marine Research Projects
EDMERP

Data

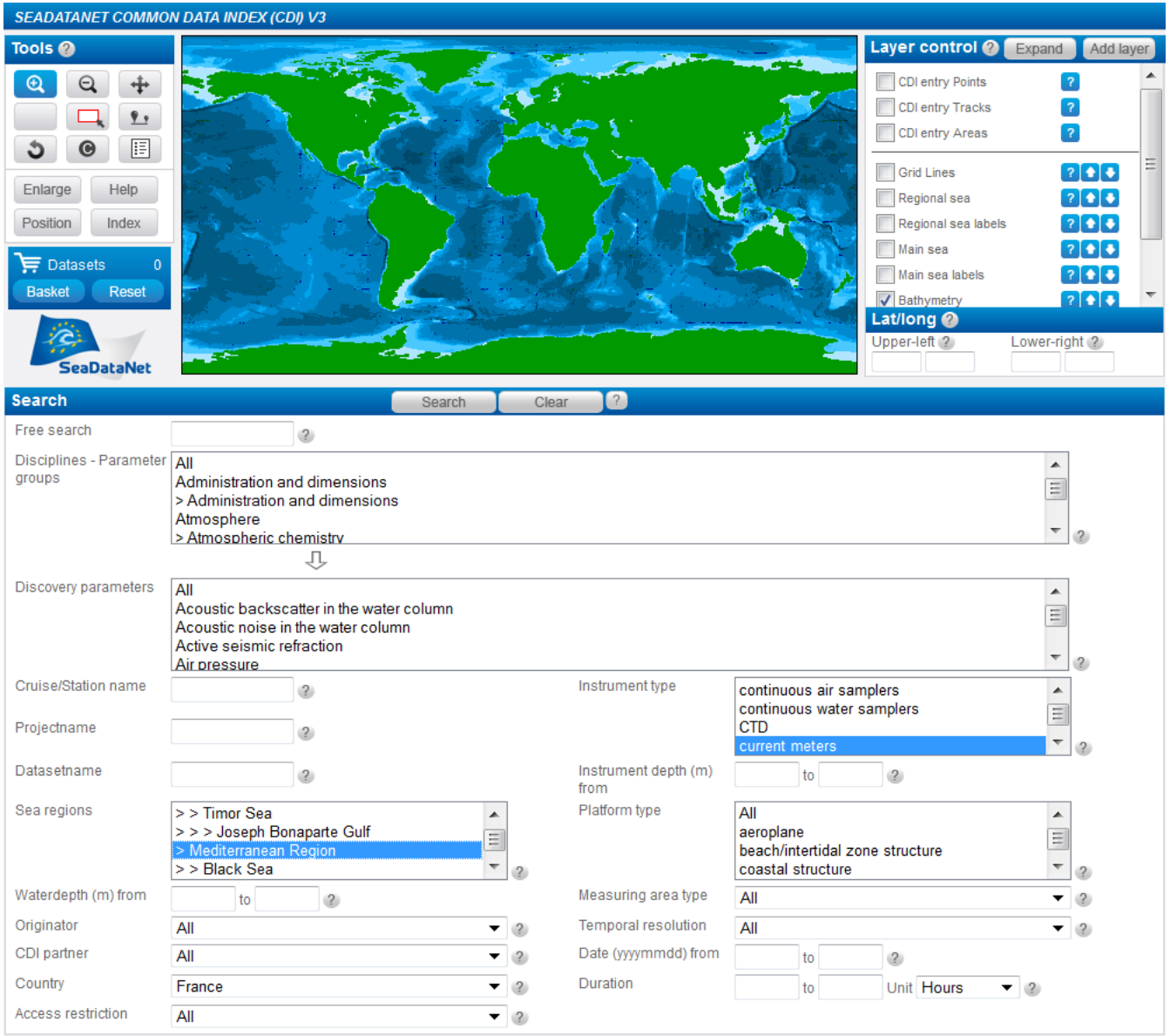


Common Data Index *Detailed Description*
CDI

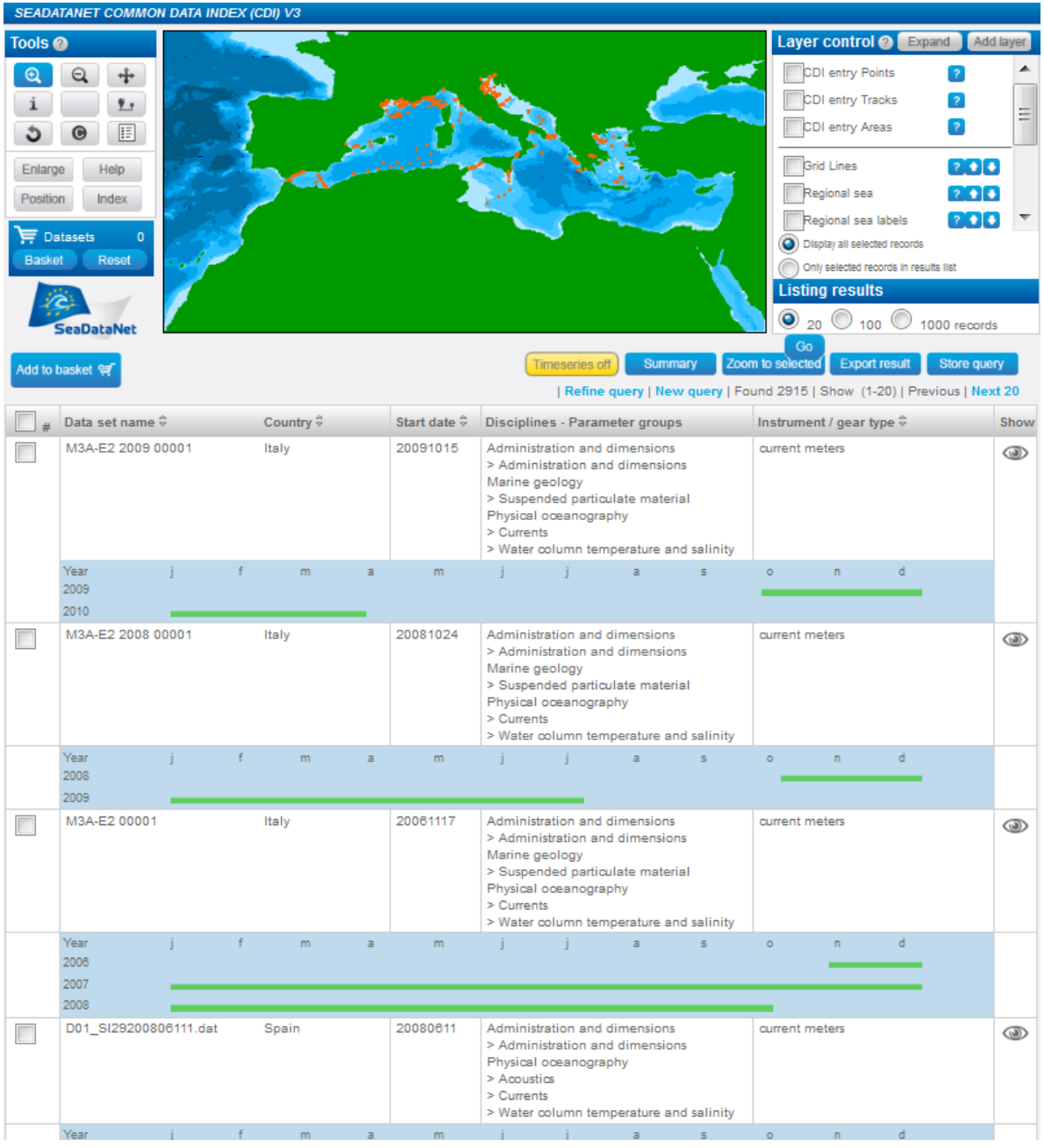
Products



Requête


SEADATANET COMMON DATA INDEX (CDI) V3
Tools
Enlarge
Help
Position
Index
Datasets 0
Basket
Reset
SeaDataNet
Layer control
Expand
Add layer
CDI entry Points
CDI entry Tracks
CDI entry Areas
Grid Lines
Regional sea
Regional sea labels
Main sea
Main sea labels
Bathymetry
Lat/long
Upper-left
Lower-right
Search
Search
Clear
Free search
Disciplines - Parameter groups
All
Administration and dimensions
> Administration and dimensions
Atmosphere
> Atmospheric chemistry
Discovery parameters
All
Acoustic backscatter in the water column
Acoustic noise in the water column
Active seismic refraction
Air pressure
Cruise/Station name
Projectname
Datasetname
Sea regions
> > Timor Sea
> > > Joseph Bonaparte Gulf
> Mediterranean Region
> > Black Sea
Waterdepth (m) from
to
Originator
All
CDI partner
All
Country
France
Access restriction
All
Instrument type
continuous air samplers
continuous water samplers
CTD
current meters
Instrument depth (m) from
to
Platform type
All
aeroplane
beach/intertidal zone structure
coastal structure
Measuring area type
All
Temporal resolution
All
Date (yyyymmdd) from
to
Duration
to
Unit
Hours

Résultats (cas des séries temporelles)



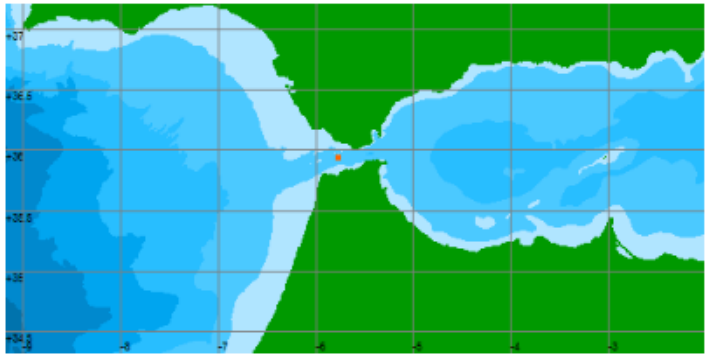
Métadonnées

Details

XML

WHAT?	
Data set name	D01_SI29200806111.dat
Discipline	Administration and dimensions Physical oceanography
Category	Acoustics Administration and dimensions Currents Water column temperature and salinity
Disciplines - Parameter groups	Date and time Horizontal velocity of the water column (currents) Sound velocity and travel time in the water column Temperature of the water column Vertical spatial coordinates Vertical velocity of the water column (currents)
GEMET-INSPIRE themes	Oceanographic geographical features
Abstract	current meter time series
Data format	MEDATLAS ASCII **Version** 2.0

HOW?	
Instrument / gear type	current meters
Platform type	mooring
Cruise name	INGRES 0308
Alternative cruise name	SI29200806111
Cruise start date	20080611
Station name	SI2920080611103080
Alternative station name	SI2920080611103080
Station start date	20080611

WHERE?	
Map	
Latitude 1	35.9113
Longitude 1	-5.7442
Datum	World Geodetic System 84
Measuring area type	point
Water depth (m)	290
Depth reference	sea level
Minimum instrument depth (m)	284
Maximum instrument depth (m)	284
Sea regions	Strait of Gibraltar

WHEN?	
Start date	20080611
Start time	13:00:00
End date	20080626
End time	17:00:00

WHO?	
Originator	Malaga University (UMA). Applied Physics departament II
Data Holding centre	IEO/Spanish Oceanographic Institute
Project name	Water exchanges through the Strait of Gibraltar and their response to meteorological and climate forcing

HOW TO GET THE DATA?	
Data Distributor	IEO/Spanish Oceanographic Institute
Access/ordering of data	web data access with registration
Internet access/ordering	
Access restriction	by negotiation

Rapport d'étape

- **Un système qui répond aux objectifs**
  - Gestion partenariale des données
  - Visibilité des producteurs de données
  - Possibilité pour l'utilisateur d'accéder à des jeux harmonisés

- **Une mise en œuvre réaliste**
  - Des coûts de mise en œuvre maîtrisés pour les fournisseurs de données
  - Une mutualisation possible avec l'implémentation de la directive INSPIRE
  - Nécessite cependant :
    - accès à des compétences techniques informatiques : mise en place, surveillance de bon fonctionnement
    - une organisation préalable de la gestion des données
      (constitution des métadonnées, harmonisation (qualité, formats...), suivi des requêtes si données protégées

- **Des limitations et de nécessaires évolutions**
  - SeaDataNet : portail web et performances d'accès limitent actuellement l'utilisation du système
    - Pas de données « temps réel », une architecture centralisée de type Coriolis reste nécessaire
    - l'utilisateur est un humain, pas d'accès de machine à machine
  - Big « metadata » : Comment générer facilement et utiliser pleinement les métadonnées?
  - Big « data » : Comment faciliter l'utilisation de larges collections de données?
    - Le volume n'est pas le seul facteur de complexité!
    - Hétérogénéité,
    - Nombre de fichiers, ...

Génération des métadonnées

- **Les métadonnées : un « mal » nécessaire**
  - Identifier les conditions d'observation (visibilité, ...)
  - Systèmes automatisés :

    augmentation du nombre de capteurs,
    chaine de production de l'observation allongée (transmission, décodage, ...)
  - A relier avec la supervision de réseau de capteurs

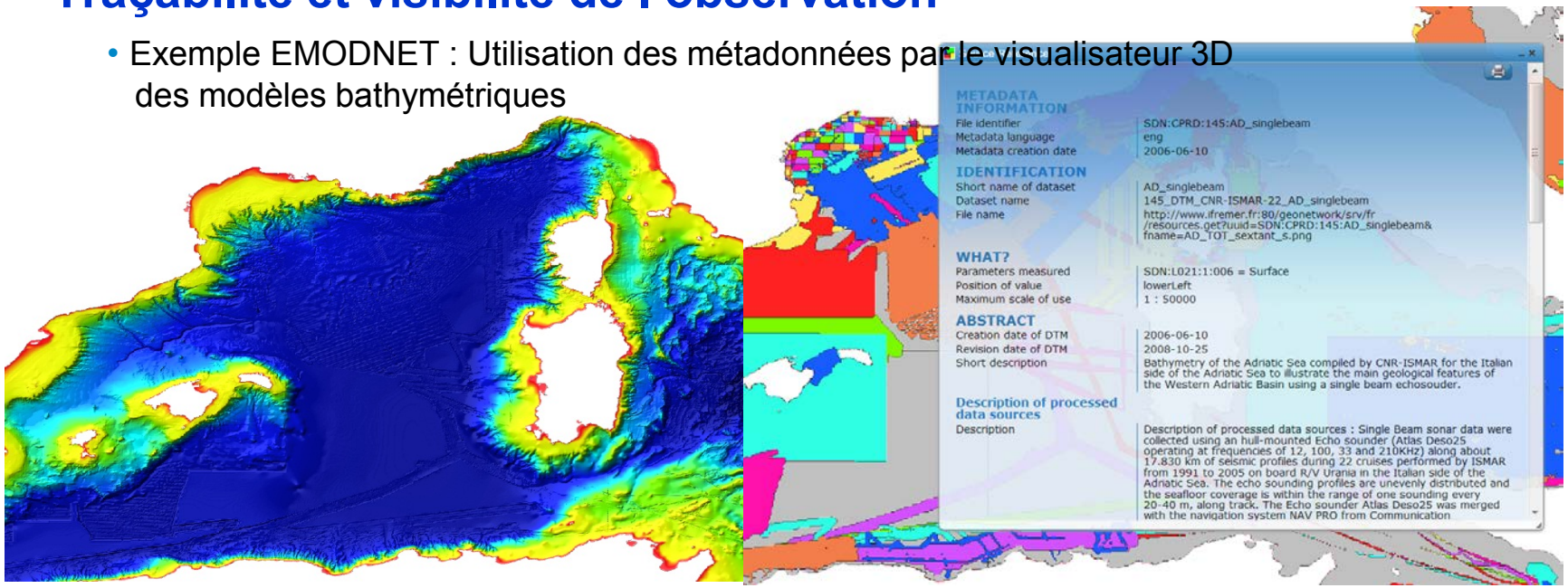
- **Conditions de génération**
  - Métadonnées générées au plus proche de l'observation pour être pertinentes
  - Eviter au maximum la saisie (effort important, risque d'erreurs)
  - Idée de « Smart Sensors » capables de
    - S'identifier
    - Décliner ses caractéristiques techniques (plage d'utilisation, étalonnage, ...)
  - Les normes existent (SensorML par exemple)

➔ **Convaincre les producteurs d'instruments**

Utilisation des métadonnées

• Traçabilité et visibilité de l'observation

- Exemple EMODNET : Utilisation des métadonnées par le visualisateur 3D des modèles bathymétriques



➔ **Permanence des liens entre vocabulaire, métadonnées et données (stabilité des URL) ... sans figer le système**

➔ **Web « sémantique » : permettre à l'utilisateur de hiérarchiser les liens**

• Une interface entre Données et publications scientifiques

- Adoption des DOI (Digital Object Identifier) pour la citation des jeux de données (avec Data Cite, opéré en France par une filiale du CNRS)

➔ **Stabilité des jeux de données**

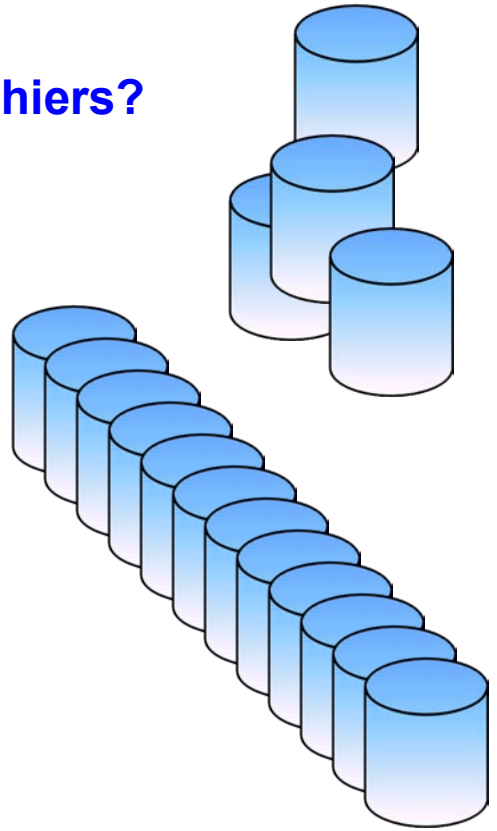
➔ **Choix de la granularité des jeux de données**

Big data ...

- **Exemple de SeaDataNet :**
  - les données élémentaires sont de petit volume (< 1Mo)
  - les données élémentaires sont organisées en fichiers
    elle viennent de plusieurs fournisseurs de manière asynchrone
  - le nombre de fichiers fait problème (centaines de milliers)
    - Limites des Operating Systems standards (nombre de fichiers par répertoire)

- **Quelle organisation pour fournir ces ensemble de fichiers?**
  - Répartir les fichiers en répertoire
    - Type d'instruments
    - Date d'aquisition
    - ???
  - Accoler un index avec les ensembles de fichiers
    - Tableau d'index (.csv)
    - Thredds Catalogue (Fichier XML structuré)
    - Data Collection (SeadataNet / Ocean Data View)

➔ **Pas d'organisation optimale générique....**



En guise de conclusion

- **Priorité sur:**
  - Préservation à long terme des données

    autour de 50% des données perdues en 10 ans si non archivées
  - Gouvernance et procédures harmonisées
    - Vocabulaire, descriptifs (métadonnées) communs, niveau de qualité, définition des droits d'accès ...
    - Modèles de données formalisés (objets métiers) : profils verticaux, séries temporelles, ...
  - Services communs
    - Basé sur les standards de gestion de l'information géo-référencée, en conformité avec la directive Inspire
- **Peu d'investissement sur les techniques informatiques nouvelles**
  - Technologies informatiques « nouvelles » (cloud, grilles, ...)
  - Méthodes d'auto-découverte (moteurs de recherche, extraction de connaissance) qui nous paraissaient difficiles à mettre en œuvre sur des données numériques hétérogènes
- **De nouvelles perspectives**
  - Une ouverture internationale (projet ODIP)
    - Etats-Unis (NOAA, Universités, Instituts de Recherche, programmes), Canada, Australie
  - Une démarche similaire en France
    - Gestion partenariale des données marines au sein d'un Pôle Océan (ALLENVI)

➔ **Résoudre le problème de complexité par :**

➔ **Des procédures harmonisées, définies à priori, pour la gestion des données,**

➔ **La distribution du travail auprès des différentes bases.**

Merci de votre attention.

http://www.seadatanet.org/
http://www.geo-seas.eu/

http://www.emodnet-hydrography.eu/
http://www.emodnet-chemistry.eu/

http://www.odip.eu/