DATARMOR : UNE INFRASTRUCTURE INFORMATIQUE POUR LES SCIENCES MARINES

TERATEC 12 JUIN 2019

Spécialisation « sciences marines »

Historique d'Ifremer et des partenaires

L'océan : un milieu partagé (et en mouvement)

Centre de données SISMER :

- certifié au niveau UNESCO/IODE
- en attente de certification RDA « Core Trustworthy repository »
- labellisé au niveau national « IR Data Terra » (ex Pôles de données et de services du système Terre)
- leader/partenaire de nombreux projets européens

Evolution du contexte

Augmentation régulière du volume des données due à :

- L'automatisation des systèmes d'observation
- Les progrès technologiques sur les capteurs
- La récupération de données d'opportunité

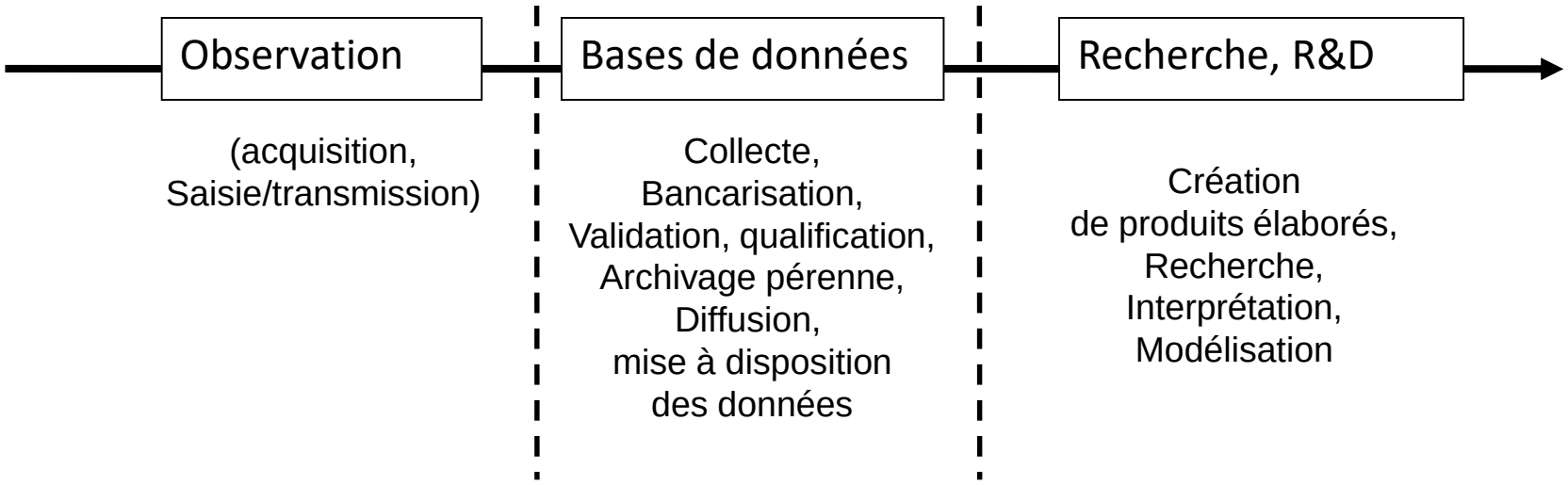
➢ Pour certains traitements, les accès disque sont devenus le facteur limitant

Les projets de recherche

Utilisateurs Ifremer, ou partenaires Datarmor ou partenaires UMR.

- Les sciences marines sont très « data driven » et ont souvent une approche « écosystémique »
- Les quelques modélisateurs en phase de recherche algorithmique disposent de comptes sur les centres nationaux
- Quelques sous-domaines particuliers :
  - **Bio-informatique**
  - **Océano spatiale**

IFREMER : Les bases de données marines



Phase « observation » :

Principaux producteurs de données

TGIR FOF : Flotte Océanographique Française

TGIR Euro-Argo (et projet inter-organisme Coriolis)

SIH (Système d'Information Halieutique)

Réseaux d'observation pour la surveillance de l'environnement et du littoral

Missions satellite intéressant l'océan

IFREMER : Les bases de données marines

Phase « bases de données » :

Organisation par filières

Données et méta-données

Les méta-données répondent notamment aux questions : Quoi? Où? Quand? Comment?

Traçabilité : nécessité scientifique et reconnaissance du producteur

Phase « Recherche, R&D » :

Nombreuses et anciennes collaborations sur les sciences marines aux niveaux national et international

La mise en commun des données marines est une nécessité scientifique et technique (contraintes/volumes parfois importants)

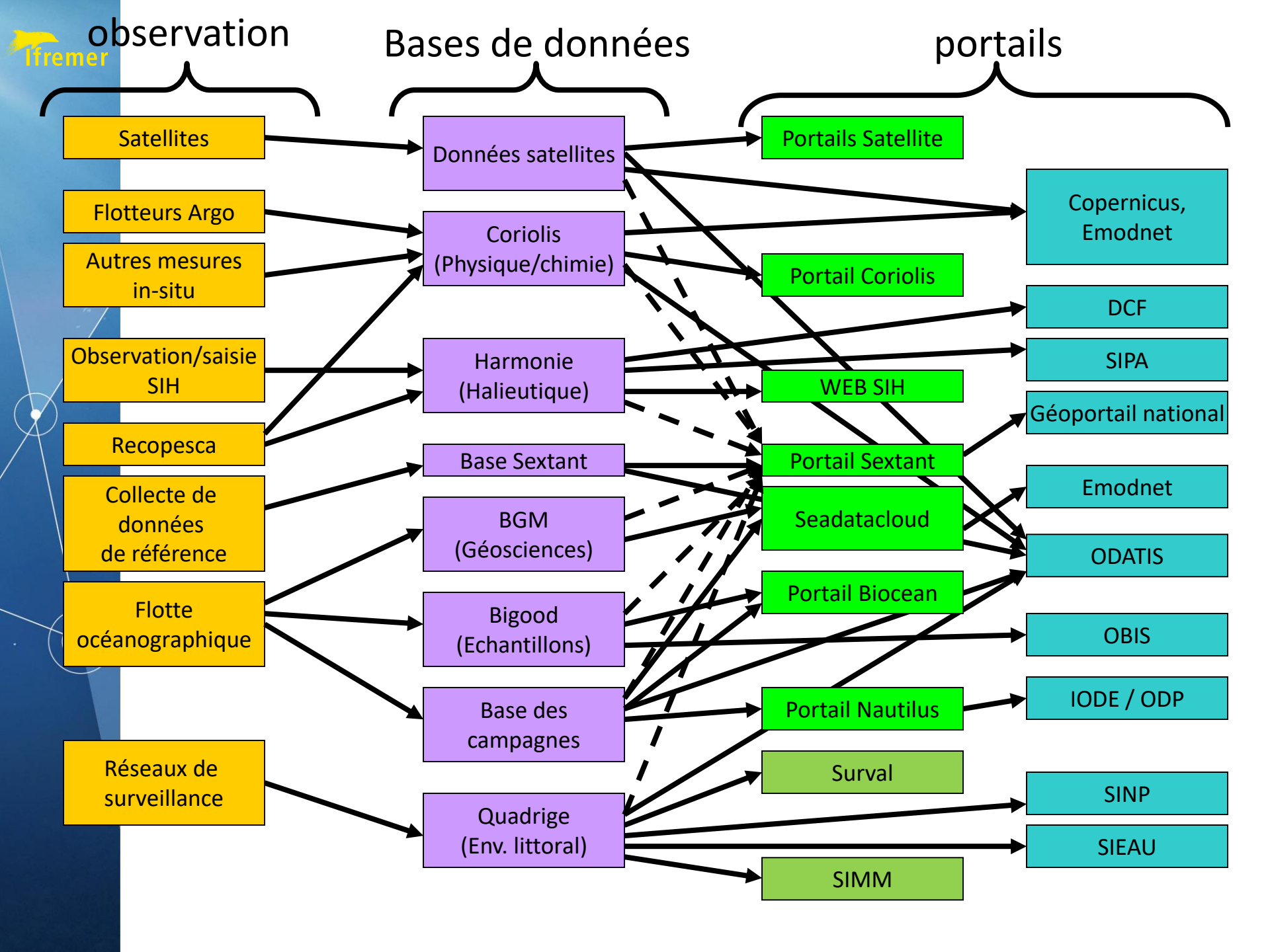
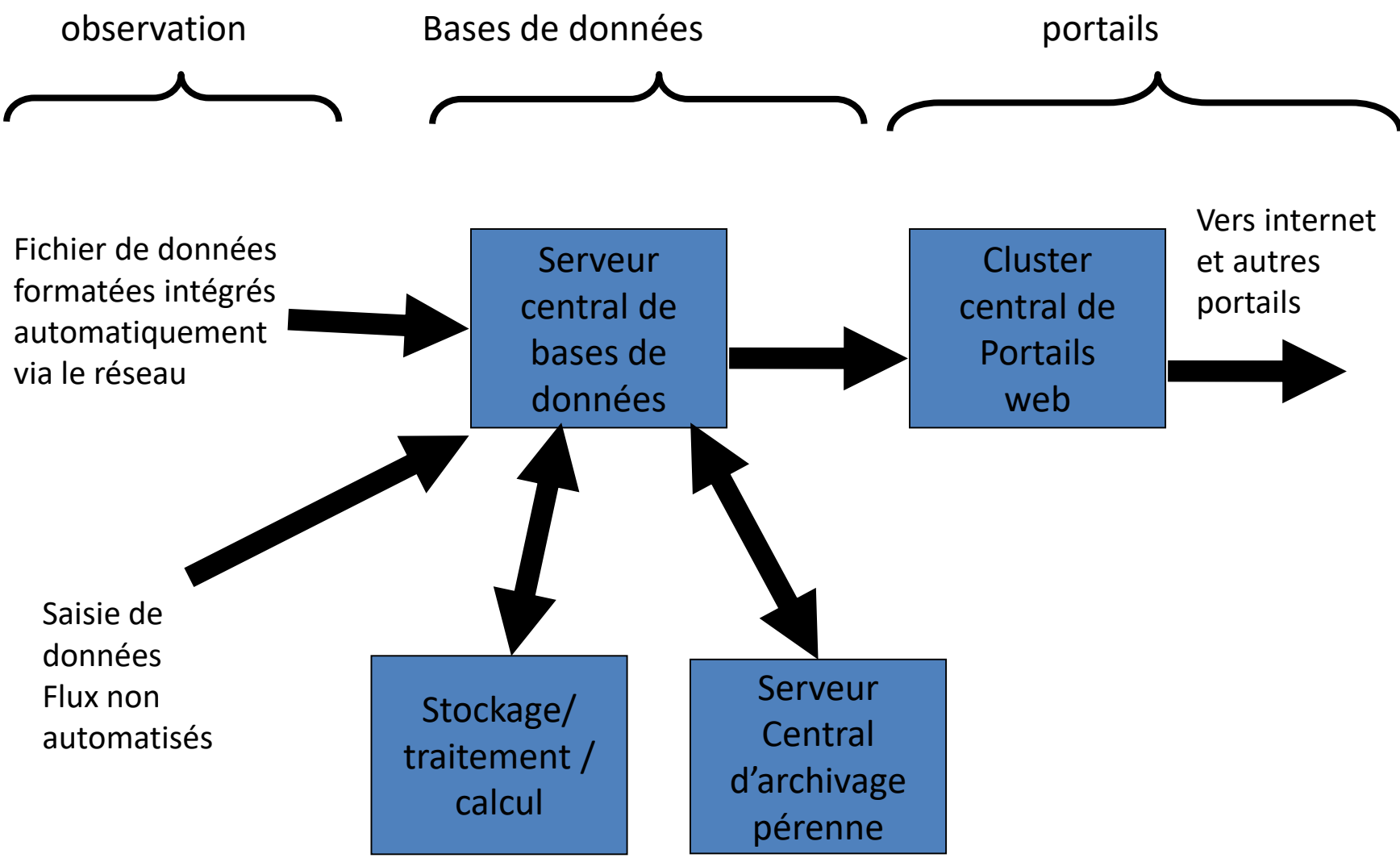

observation
Bases de données
portails
Satellites
Flotteurs Argo
Autres mesures in-situ
Observation/saisie SIH
Recopesca
Collecte de données de référence
Flotte océanographique
Réseaux de surveillance
Données satellites
Coriolis (Physique/chimie)
Harmonie (Halieutique)
Base Sextant
BGM (Géosciences)
Bigood (Echantillons)
Base des campagnes
Quadrige (Env. littoral)
Portails Satellite
Portail Coriolis
WEB SIH
Portail Sextant
Seadatacloud
Portail Biocean
Portail Nautilus
Surval
SIMM
Copernicus, Emodnet
DCF
SIPA
Géoportail national
Emodnet
ODATIS
OBIS
IODE / ODP
SINP
SIEAU

Schéma au niveau informatique



Infrastructure informatique en appui au centre de données marines

Cluster de base de données relationnelle ORACLE (plus de 20 milliards d'enregistrements)

Service de sauvegarde

Service d'archivage pérenne

Infrastructure d'hébergement Web : Intra-, Extra- et Inter- net, avec annuaires associés

… et Datarmor

Datarmor : projet CPER Bretagne 2014-2020

Datarmor : conventions d'appellation

Capacité de calcul / d'exécution

- ClusterHPC : nœuds de calcul MPI
  (396 nœuds INTEL SGI ICE-XA 28 cœurs)
- ClusterSMP : nœuds de calcul à mémoire partagée
  (nœuds INTEL SGI UV3000 et UV 2000)
- ClusterWEB : nœuds orientés services Web
  (nœuds INTEL SGI et HPE)

Datarmor : conventions d'appellation

Capacités de stockage/disque :

- Scratch : espace de stockage temporaire pour les nœuds de calcul (0,5 Pio DDN sous LUSTRE)
- DataWork : espace de stockage « de travail » (8 Pio DDN sous GPFS)
- DataHome : espaces perso (40 Tio NFS sauvegardé)
- DataRef : espace de stockage pour données « de référence » (1,5 Pio DDN sous LUSTRE)
- DataWeb : espace de cache applicatif propre au ClusterWEB (100 Tio NFS sans sauvegarde)

La problématique du partage des « ressources »

Traitements batchs (facile) :

- Réservation exclusive mais temporaire de ressources : partage par étalement dans le temps (gestionnaire batch)

Traitements interactifs (Web, et autres) :

- Pas de réservation exclusive : partage statistique sans garantie de temps de réponse (usage de la virtualisation)

Stockage :

- Réservation exclusive (et quasi-définitive). Doit faire l'objet d'extensions pour répondre aux nouvelles demandes

Les mécanismes de gouvernance

Les infrastructures informatiques se déclinent en unités d'œuvres, avec chacune des coûts de revient. Exemples :

- 1 Tio de stockage disque pendant 5 ans (hors sauvegarde)
- 1 millier d'heures de calcul
- 1 Tio d'archivage pérenne (ad vitam aeternam)
- 1 million d'enregistrements dans la BDD de catalogage
- 1 compte informatique pendant 1 an

Les filières de données en continu, les projets ponctuels et les activités scientifiques se déclinent et se quantifient en unités d'œuvre

→ Les ressources informatiques sont gouvernées par les activités thématiques

Evolution des coûts de revient des ressources informatiques

Sur la base des coûts des unités d'œuvre Ifremer :

- En 2011, pour le prix de 1 To de stockage pendant 1 an, on pouvait faire 6873 heures de calcul
- En 2019, pour le prix de 1 To pendant 1an, on peut faire 16026 heures de calcul, soit 2,33 fois plus

Il semble que cette tendance soit de « long terme »

Et pourtant 1 To de stockage de 2019, c'est beaucoup moins bien que la même chose en 2011

Alors que 1h de calcul en 2019, c'est mieux qu'en 2011

Evolution des coûts de revient des ressources informatiques

Pour répartir les ressources informatiques de manière pertinente entre les utilisateurs, il faut se placer au niveau de la thématique commune « sciences marines » :

- On peut provoquer la mutualisation de données d'intérêt commun (pour économiser de l'espace de stockage) quitte à provoquer plus de temps de calcul
  - **Exemples : résultats de modèles océano ou météo téléchargés automatiquement et quotidiennement, en configuration exhaustive, pour éviter les téléchargements et stockages individuels**
  - **Stockage de certaines données de bio-informatiques en formats compressés**
- On peut réaliser des traitements automatiques et périodiques pour gagner en performance
  - **Exemple : Cache pré-calculé pour Appli Web géographique (« tuilage »)**

Quelques autres tendances lourdes en sciences marines

Moins d'usage « académique » des langages, de la parallélisation à façon en HPC

- Les utilisateurs scientifiques maquettent sur leur « poste de travail » avec des outils de base (Matlab, R, Python, …) et essayent d'appliquer leur traitement sur le serveur central avec de gros volumes de données.
- Logiciels et pipelines « métier » pré-existants : bio-informatique, traitements acoustiques, halieutique

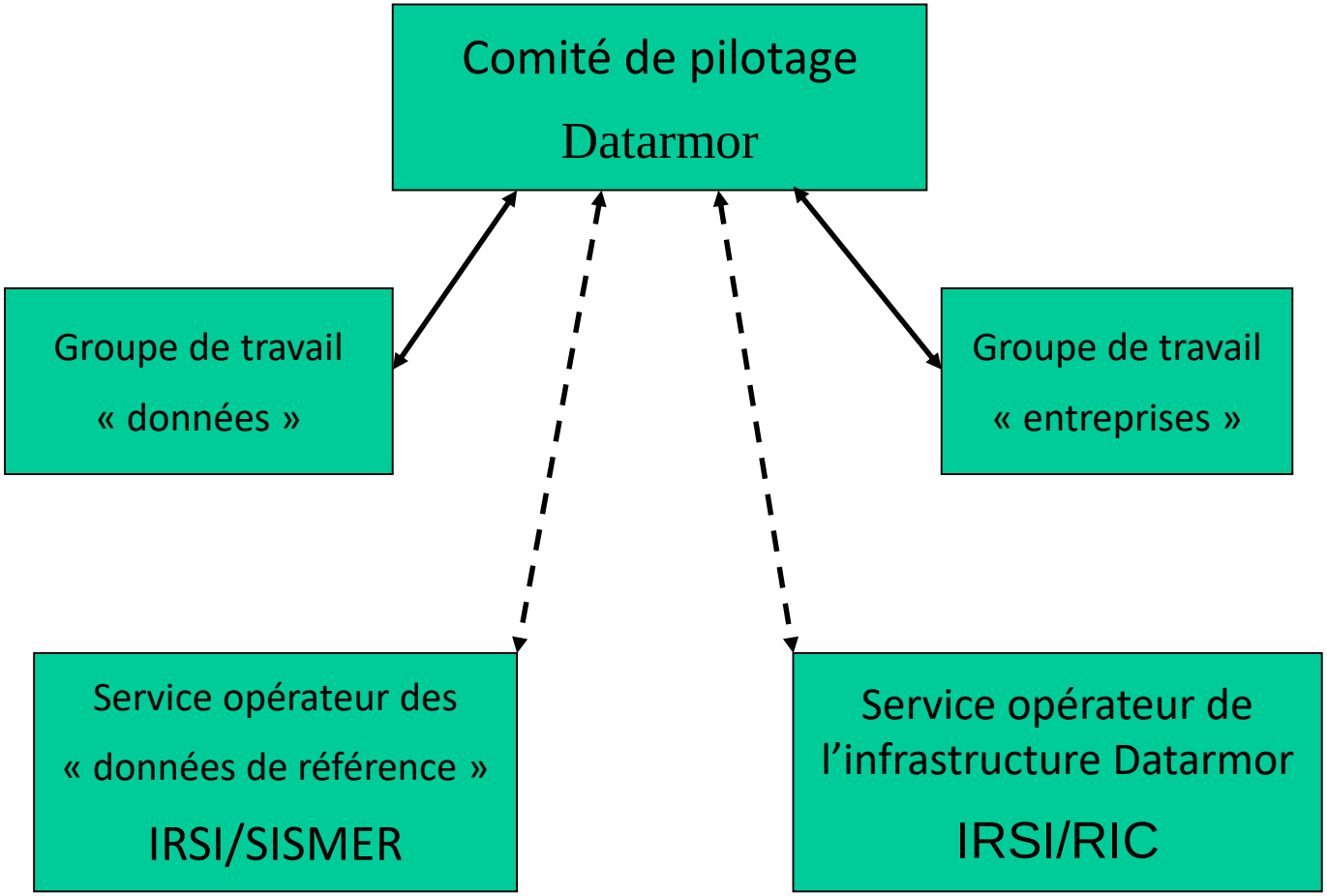
Davantage de besoins en mode interactif

- Environnements virtuels de recherche (VRE) dans le cas de projets européens ciblant l'interopérabilité

Prise en compte plus fréquente de la dimension « temps » dans les données marines

- Traditionnellement campées sur la 3D de l'espace géographique, les sciences marines évoluent vers l'étude des impacts du changement global. Les formats de fichiers de données traditionnels (beaucoup de Netcdf) ne suffisent plus.

Gouvernance



Répartition des rôles

GT données

- **Demande d'hébergement de données de référence**
- **Demande d'espace de travail partagé au titre d'une unité, d'un projet**

GT entreprises

- **Demande de labellisation d'une entreprise**

Comité de Pilotage

- **Synthèse globale**
- **Demande de labellisation d'un partenaire public**
- **Validation de la politique de partage des ressources de la machine en général (et le calcul en particulier)**

Département IRSI (Infrastructure de Recherche et Systèmes d'Information) : ~75 permanents

Pierre Cotty : Directeur

Gilbert Maudire : Directeur adjoint et Directeur de ODATIS

- Service RIC (Ressources Informatiques et Communication) : 20 permanents dont 12 administrent les ressources d'informatique scientifique
- Service SISMER (Systèmes d'Information Scientifiques pour la MER) : 25 permanents pour opérer les SI de données marines
- Service ISI (Ingénierie des Systèmes d'Information) : 15 permanents en charge des développements/projets sur les SI opérationnels
- Service de bio-informatique : 4 permanents en charge des données et applis de bio-informatique
- Service de coordination Euro-Argo : 2 permanents articulés avec l'ERIC Euro-Argo et la JCOMMOPS
- Service d'Informatique de Gestion : 6 permanents

Merci