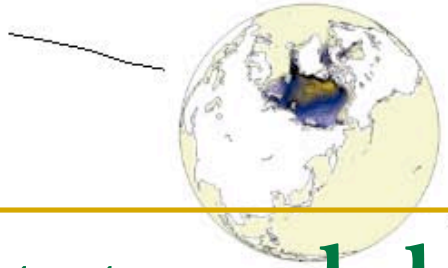
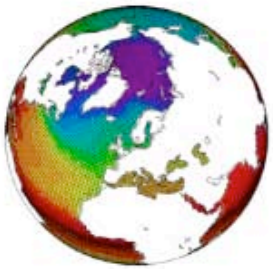
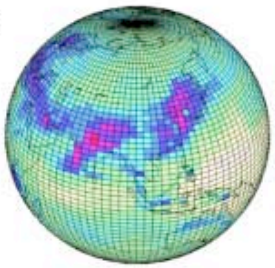


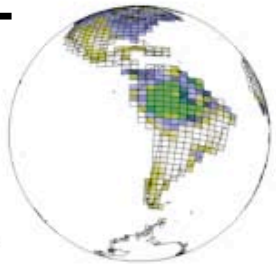
PRODIGUER un nœud de distribution des données CMIP5 GIEC/IPCC

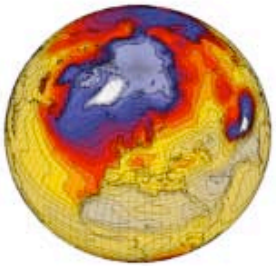


Sébastien Denvil

Pôle de Modélisation, IPSL







Institut Pierre Simon Laplace

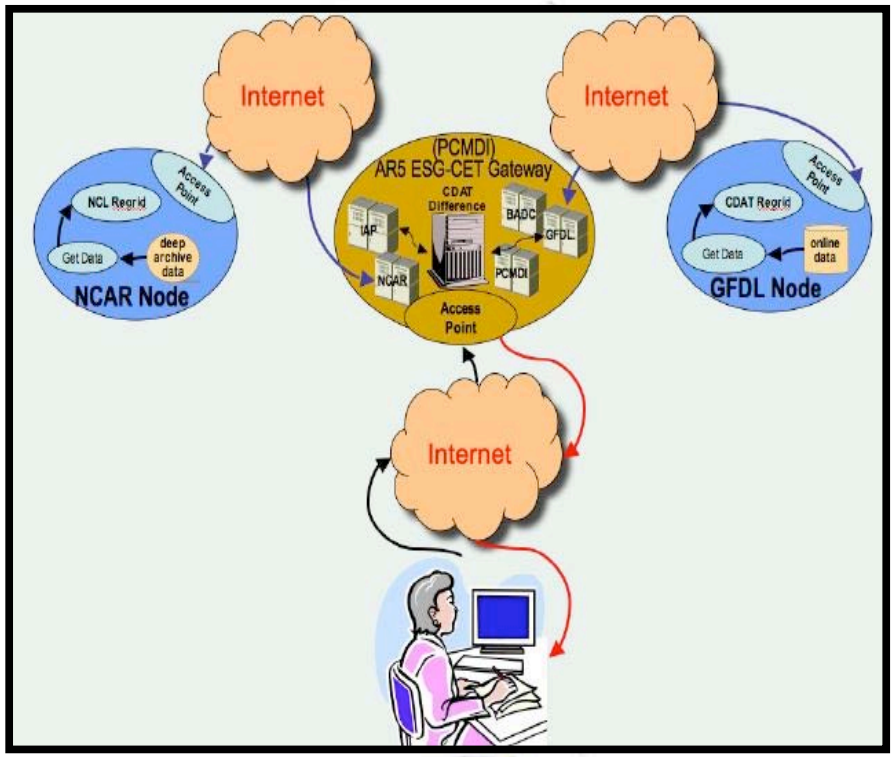
IGCMG - IPSL Global Climate Modelling Group

En route pour CMIP5?
Coupled Model Intercomparison Project (5)

- Coordination de l'aspect données :
- PCMDI: Communauté CMIP5
- BADC et WDCC: Communauté Climat Européenne et IPCC working group 2 et 3 ("mandat" IPCC-DATA.ORG)
- 3 Décembre 2008: MoU Tripartite (PCMDI, BADC, WDCC)
- > 20 groupes de modélisation du climat
- > 50 expériences numériques
- > 86 simulations pour satisfaire aux expériences
- > 6500 ans de simulation
- > 750 TB d'archive "CORE" envisagée (peut être 1PB)
- Qui sait au juste combien de TB de simulations au total?

Earth System Grid - Center for Enabling Technologies

- Fournira une architecture capable de rendre visible des données hébergées sur des “data nodes” via des “gateways”
- Support à CMIP5 à travers des “modelling-nodes” et des “core-nodes”, le premier type hébergeant les données des groupes, et le second hébergeant le “core” data défini par CMIP5.
- Plusieurs “core-nodes” sont prévus, deux en Europe (BADC, WDCC), plusieurs aux Etats Unis et un au Japon.
- Le « core-data » défini par CMIP5 est un compromis entre les plus utiles pour l’inter comparaison et ce qui est faisable



Rappel: 2 usages différents de CMIP5 “CORE”
1) CORE experiments et
2) CORE data (de toutes les expériences: CORE, Tier1 and Tier2).

Calendrier ESG/CMIP5

2008: Design and implement core functionality:

Browse and search
Registration, Single sign-on / security
Publication,Distributed metadata
Server-side processing

Early 2009: Testbed

By early 2009 it is expected to include seven centres in the US, Europe and Japan:
Program for Climate Model Diagnosis and Intercomparison - PCMDI (U.S.),
National Centre for Atmospheric Research - NCAR (U.S.),
Geophysical Fluid Dynamics Laboratory - GFDL (U.S.),
Oak Ridge National Laboratory - ORNL (U.S.),
British Atmosphere Data Centre - BADC (U.K.),
Max Planck Institute for Meteorology - MPI (Germany),
The University of Tokyo Centre for Climate System Research (Japan).

2009: Deal with system integration issues and develop production system.

Summer 2009, hardware/software requirements will be provided to those that want to be Nodes.

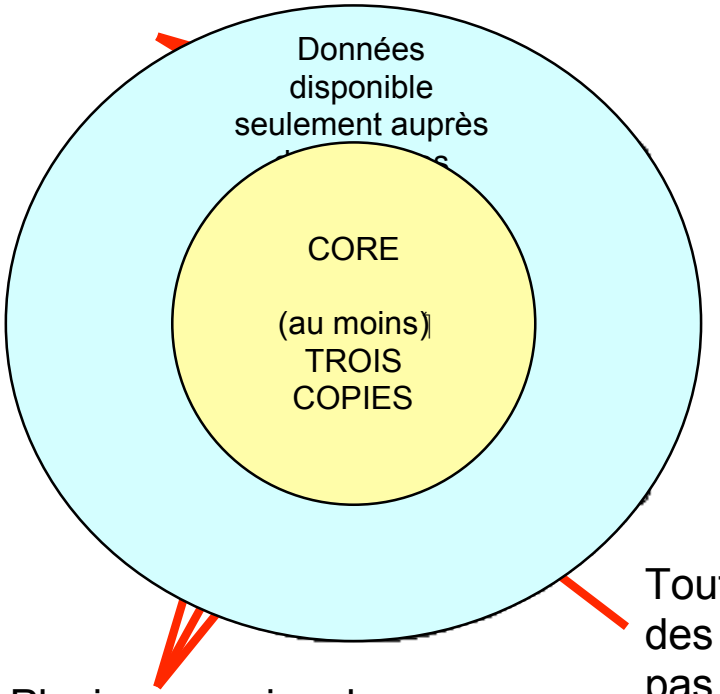
2010: Modeling centres publish data

2011-2012: Research and journal articles submissions

2013: IPCC Assessment Report #5

La Fédération Earth System Grid conduite par le PCMDI

Données disponible auprès des groupes de modélisation (via ESG), et dans de multiples “CORE centres”.



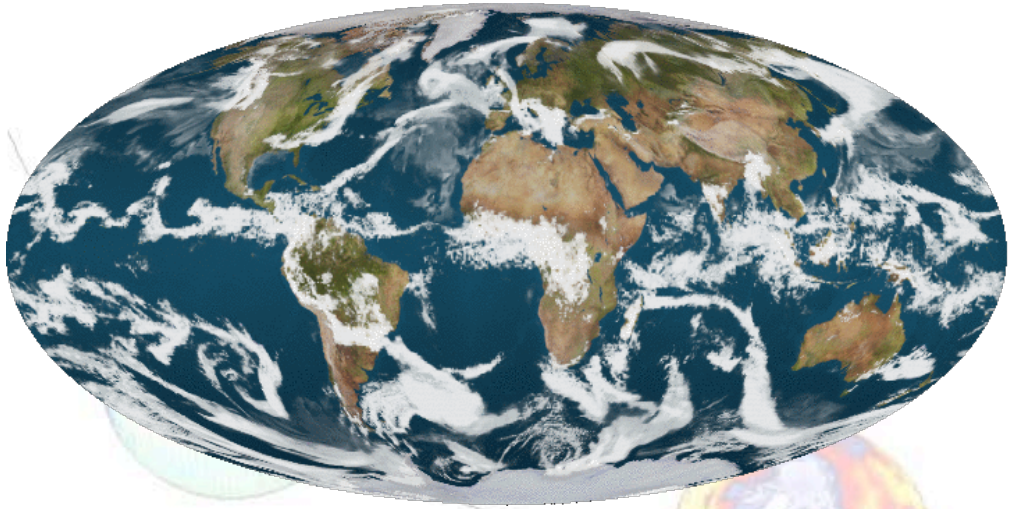
- Si chaque couleur représente les sorties d’un groupe de modélisation, alors on peut voir que la structure est en hub, avec réplication au sein des noeuds centraux, et connections à des noeuds périphériques.

Plusieurs copies du “CORE data”; ESG *et d’autres* interfaces!

Tout les centres qui fourniront des “core-data” ne rendront pas disponible le reste au sein de la fédération de sites.

Estimation du Stockage global des simulations IPSL

- Stockage brute borne basse ➔ **565 TB**
- Stockage brute borne haute ➔ **1000 TB**
- Distribution CMIP5 (25-50%) ➔ **(140-280) (250-500) TB**
- Stockage global ➔ **700-1500 TB**



LMDz 0.5° (50 Km)

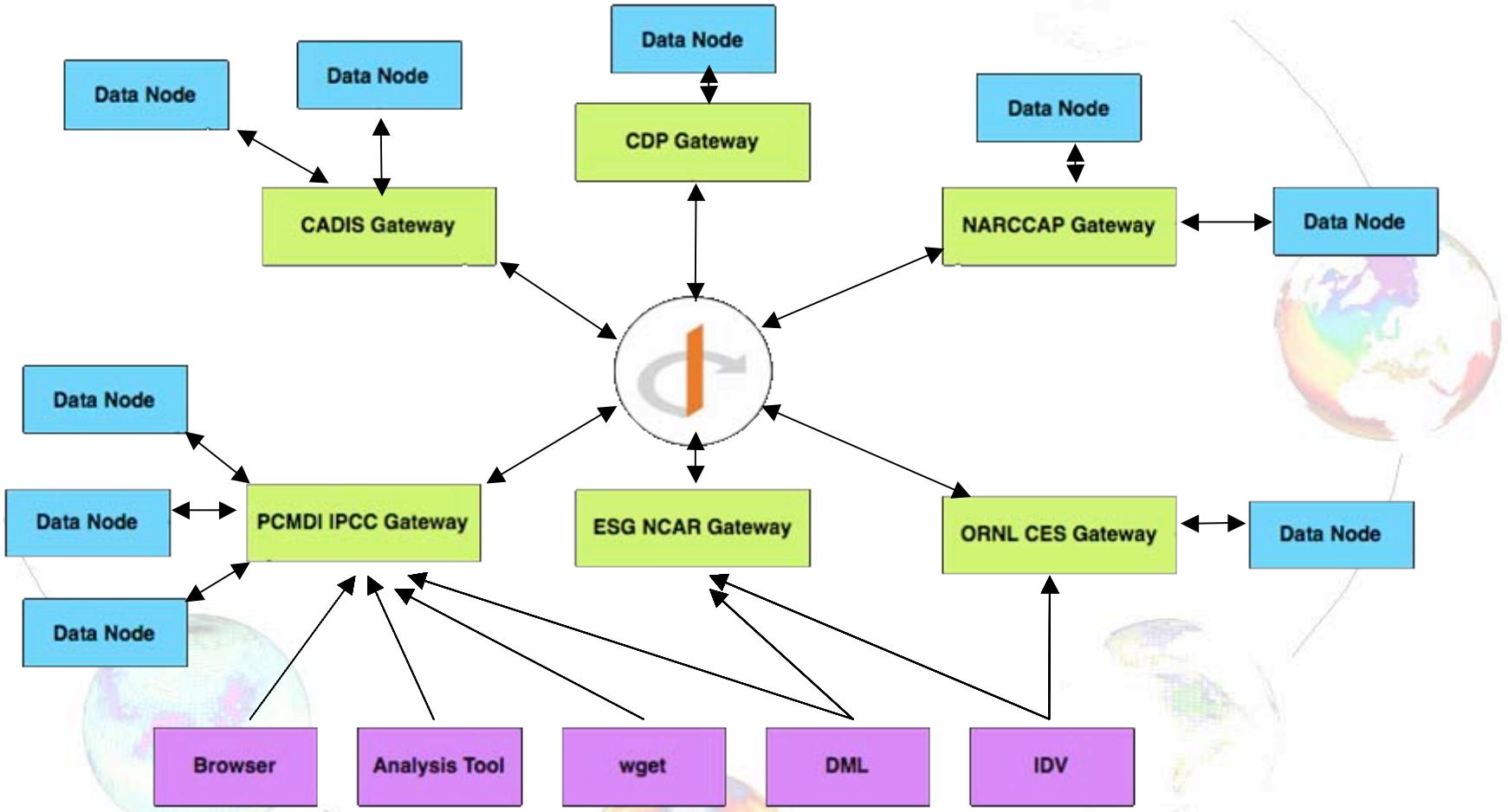
Gestion des données depuis des années

- Principalement **centralisée**, stockée sur un san
- Accès **OpenDAP** pour les centres de calcul (dods)
- Système basique de récupération de données
- Accès à des **données brutes**
- Sécurité/Authentification/Restriction d'accès aux données : pas un problème
- Pas de post-processing à la demande
- Pas d'intégration des metadata
- Ne supporte pas les **requêtes d'interrogation** de haut niveau

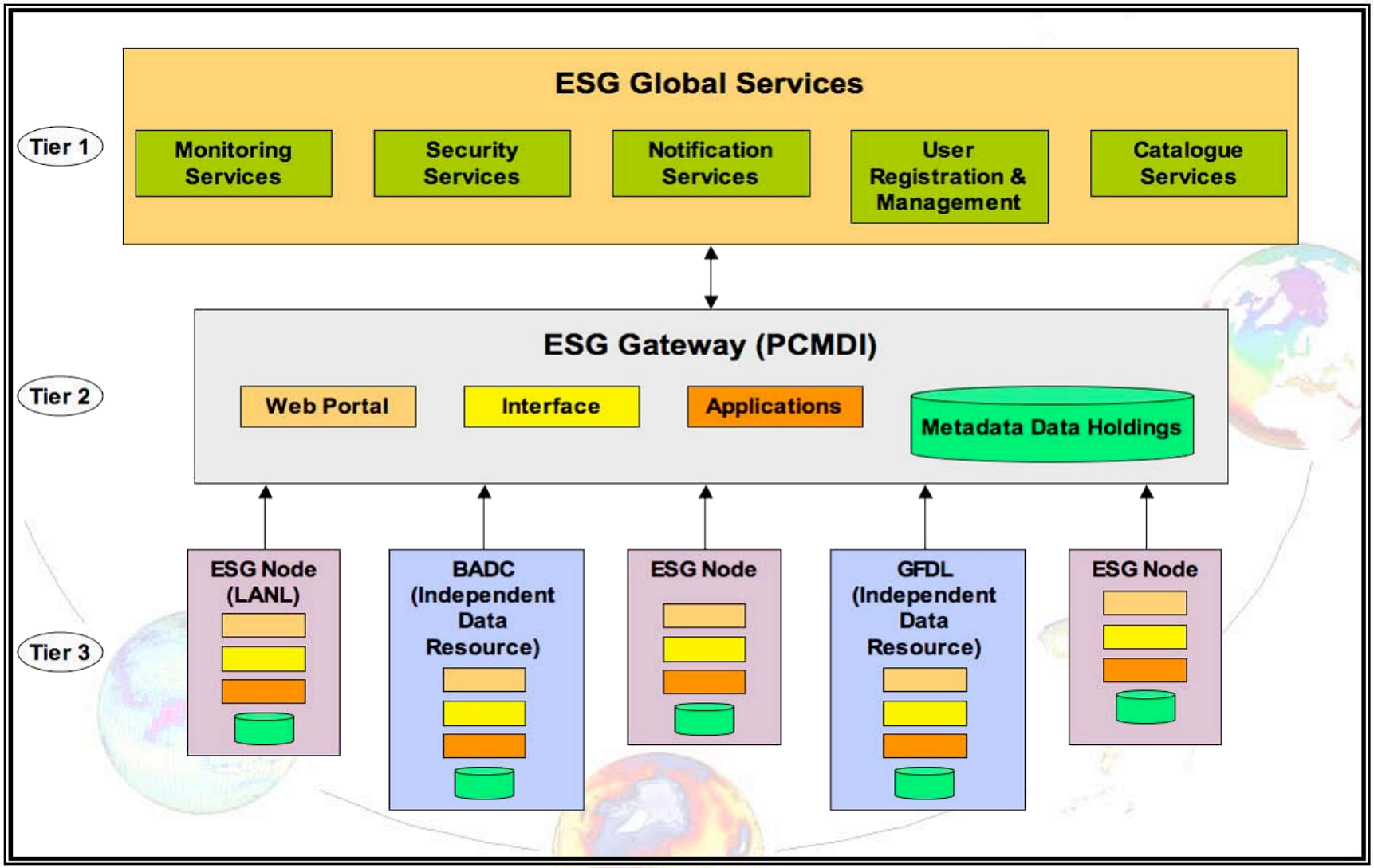
Gestion des données par Prodiguer

- Bouger les données un **minimum**, les garder proche des centres de calcul d'origine quand c'est possible
  - Protocole d'accès aux données, liens forts avec les centres de calcul
- Quand on doit bouger les données, le faire vite et avec un **minimum d'intervention humaine**
  - Management des ressources de stockage, réseaux rapides
- Garder une **trace** de ce que l'on a, de ce qui est sur "deep storage", suivi des accès
  - Metadata et Catalogues de données
- Exploiter une **fédération de sites**
  - Couche logicielle Earth System Grid

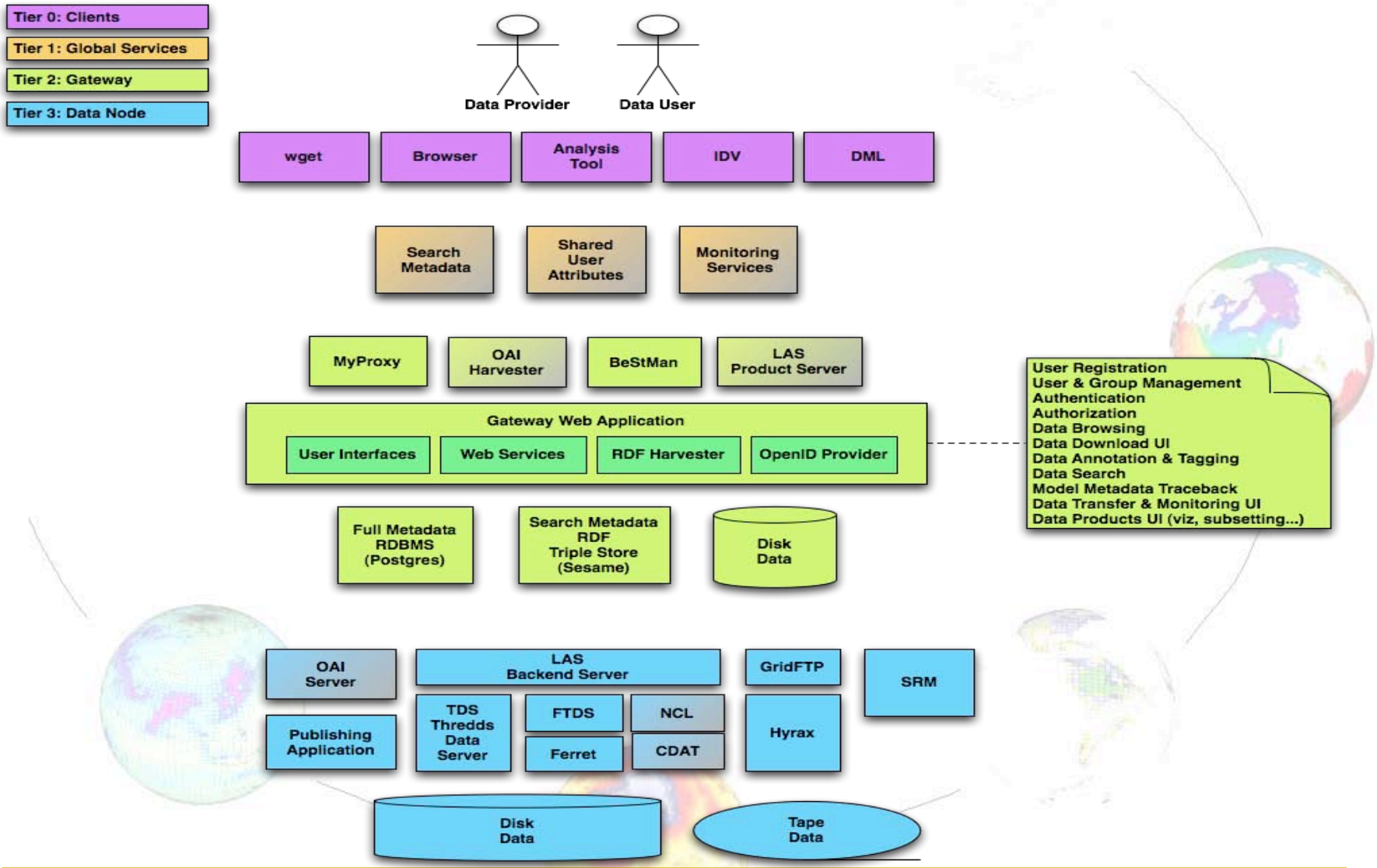
Fédération de sites Earth System Grid



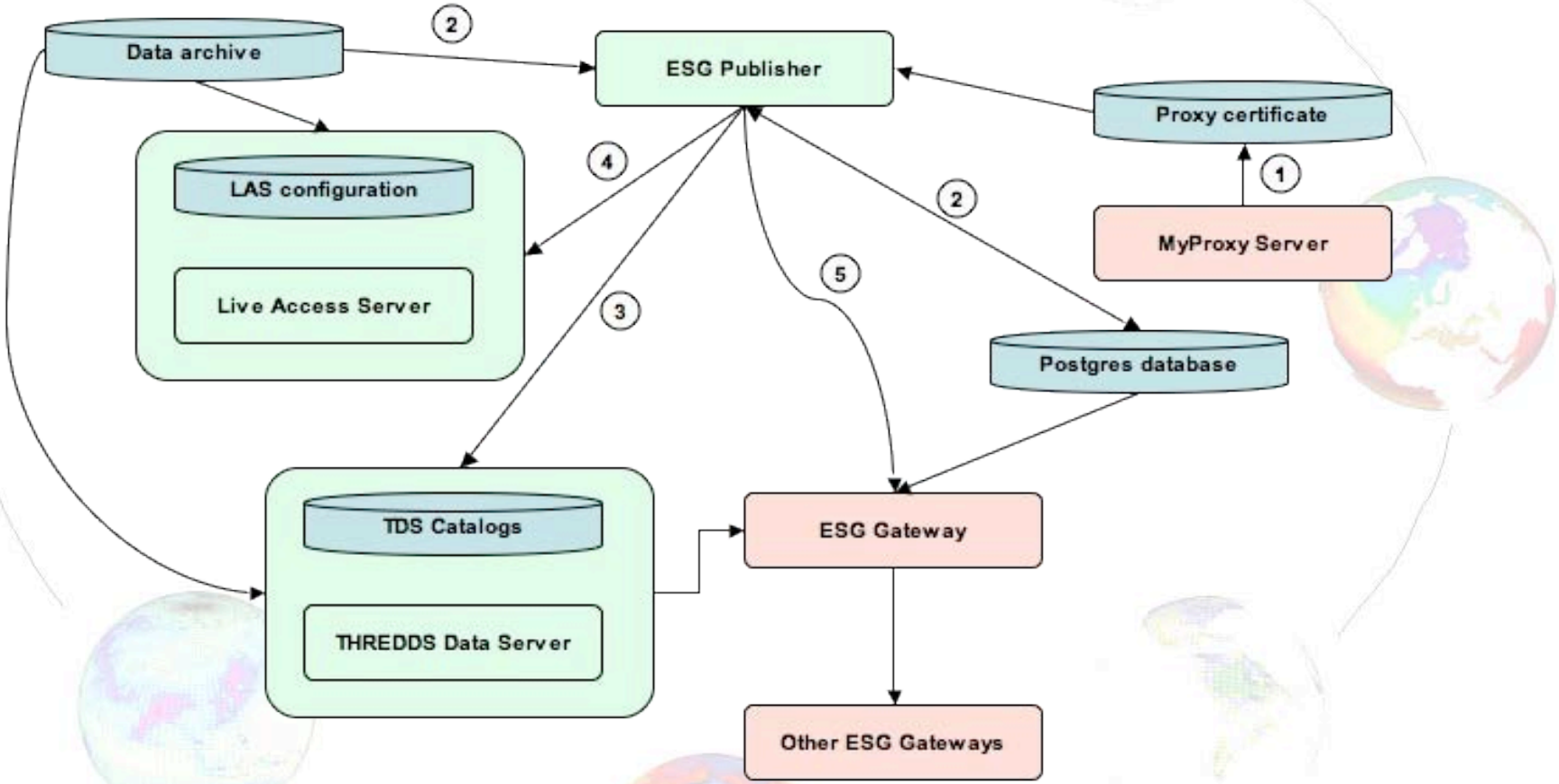
Earth System Grid en 3 Tiers.



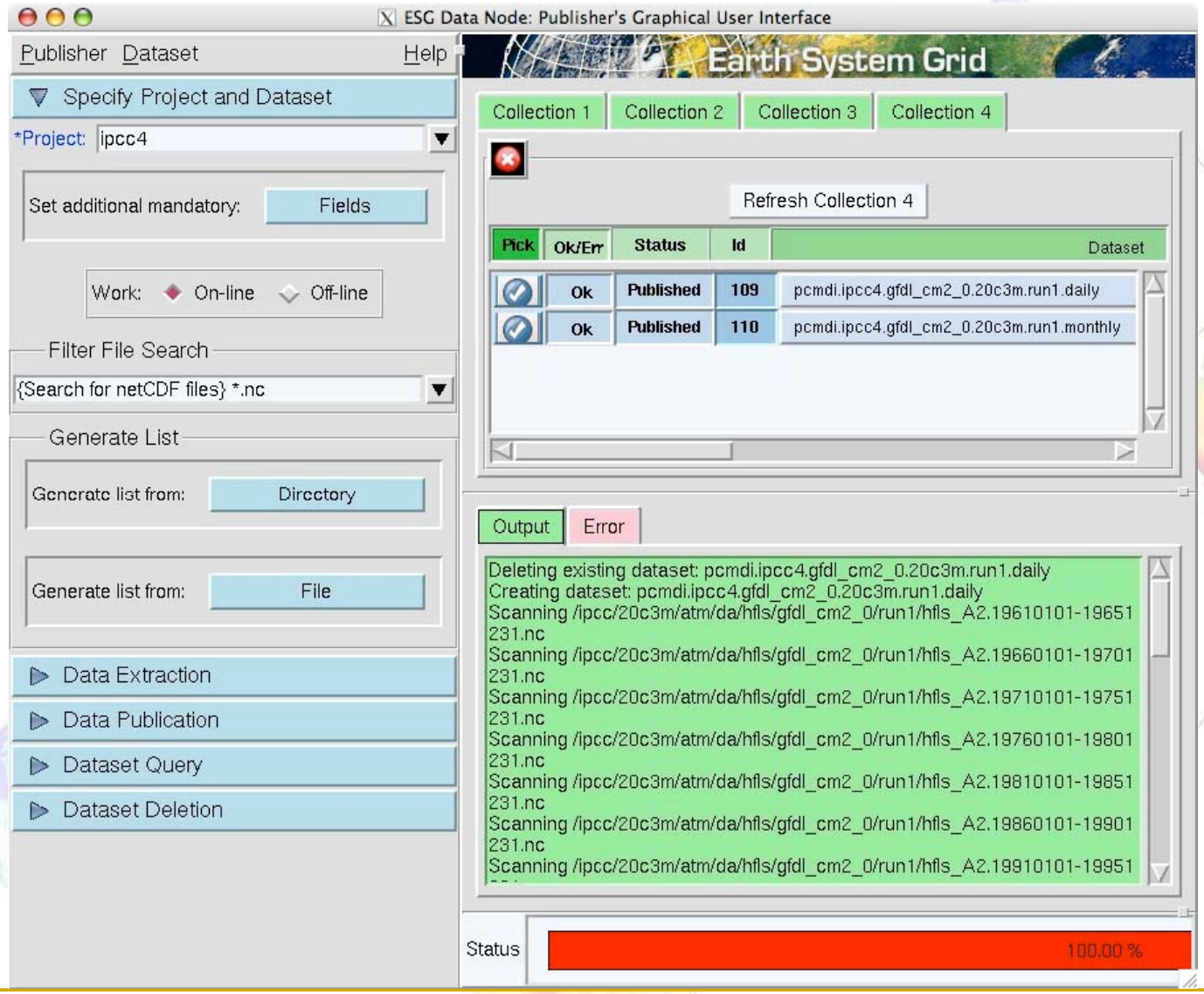
Architecture Earth System Grid



Publication des données par les “data nodes”



ESG Data Node Publisher GUI



Conclusions

- Étroite collaboration avec les centres de calculs
- Espace de stockage adéquate (cache système, disponibilité des fichiers)
- Interconnections des centres de calcul (utilisation croisée des données)
- Logiciel de Tiers3 ESG sur les centres
  - Installation couche logicielle
  - Ouverture de services

Perspectives

- La réponse Européenne à la prolifération des simulations de modèles climatique s'écrit en partenariat avec les américains du consortium ESG-CET.
- Venir en support de CMIP5 requiert un travail sur l'environnement logiciel, le stockage de données, leur manipulation, la distribution aux utilisateurs ET de décrire les simulations, leurs contextes, et les données résultats.
- Financement national (pour le stockage des données CMIP5/AR5), et le support au travail de distribution des données.
- Un certain nombre d'initiatives de financement Européenne majeure, parmi lesquelles:
- Financement CE pour développer les infrastructures:
  - METAFOR
  - IS-ENES (et d'autres financements nationaux en support d'IS-ENES).