How will we manage large multidisciplinary scientific datasets

Michel Hoepffner, Nathalie Fourès, Hassan Makhmara and Fernando Niño

(Medias-France)

Preservation versus added-value

- The best long-term preservation for scientific data: have it stored in technical centers managed by data experts
- The best added-value for scientific data: to timely update datasets by the scientists in charge

We think it's possible to solve this apparent dilemma

The actors

- Scientists:
  - Research teams, represented by their Principal Investigators
  - Users (the scientific community)
  - Multidisciplinary international scientific programs: IGBP 2, WCRP, IHDP, etc.
- Operators:
  - National agencies providing data (space agencies, etc.)
  - Technical operators like Medias-France

Medias-France in few words with:

A SERVICE STRUCTURE:

- **DATABASE AND INFORMATION SYSTEM DESIGN AND MANAGEMENT**
- PROJECTS SUPPORT (OBSERVING SYSTEMS AND NETWORKS, etc.)
- TRAINING (FELLOWSHIPS, etc.)
- CONSULTANCY AND EXPERTISE (GMES, …)

MEDIAS

http://medias.obs-mip.fr

Development and management of database and information systems

	Sea/ Atmosphere/ Hydrology/	/Atmospheric chemistry	Earth Science	Environment	Tools
Ended	EMET-CEP HAPEX-SAHEL JGOFS1 ELMASIFA EPITHERME FETCH* JGOFS2*	EXPRESSO IDAF MOZAIC PIC DU MIDI O3O AIRQUAL IGAC I & S PRE-ESCOMPTE	POLLEN WDC-A (EPD) (EDDI) FORMAT APD CD-Rom Pollen	SUD-SAHEL Images-B/OSS	WEB-Site CDROM Mediterranee MEDESERT 99 UNISPACE III GIS Grass BASS 2000 CEOS-CDROM 2000
In progress	EMERCASE GMES CATCH PLUVIOM	DEBITS PREPA1* IDAF ESCOMPTE	DIAF CLEHA RESOLVE CPC ECLIPSE	ADAM MDM OSS/LIFE AID-CCD ZA/ORME	RICAMARE GICC1 SEARCH WEB-Site ISIS AMMA POLKA IMMEDIAS EUFOREO
Planned	AMMA/Histo/Méta ProMed S2E/ARGOS IMFREX MEDWATER CHOLCLIM ENSO/PNEDC	IGAC Meta PREPA2* AMMA/Chimie	PAGES/Metadata XPROXY GPS	BIODIVALP GEOLAND/ POSTEL GMES-NOW EDEN ORE/RETYS ZA/APM	GICC2 CIESIN ACMAD IMEDIAS/ISS S2E/ARGOS PLAN BLEU POSTEL MAZURKA GMES/informations

Database

- Disciplinary:
  - Atmosphere
    - Emet
  - Hydrology
    - Catch
  - Palynology
    - APD
- Multidisciplinary:
  - Hapex-Sahel
  - Fetch
  - Amma (African Monsoon Multidisciplinary Approach)

MEDIAS

International coordination

- Member of the PAGES Data Board (Past Global Changes) with the World Data Centers of Boulder (USA) and Bremen (Germany)
- Mirror site of the World Data Center of Boulder (USA)
- Coordinator of various European funded projects (Elmasifa, APD, Format, Search, Ricamare, etc.)

MEDIAS

TOOLS

- Management of Internet sites:
  - web
  - ftp
- Data visualization and extracting interface development
- Database network, with scientists involved in the management of data

The scientific institutions

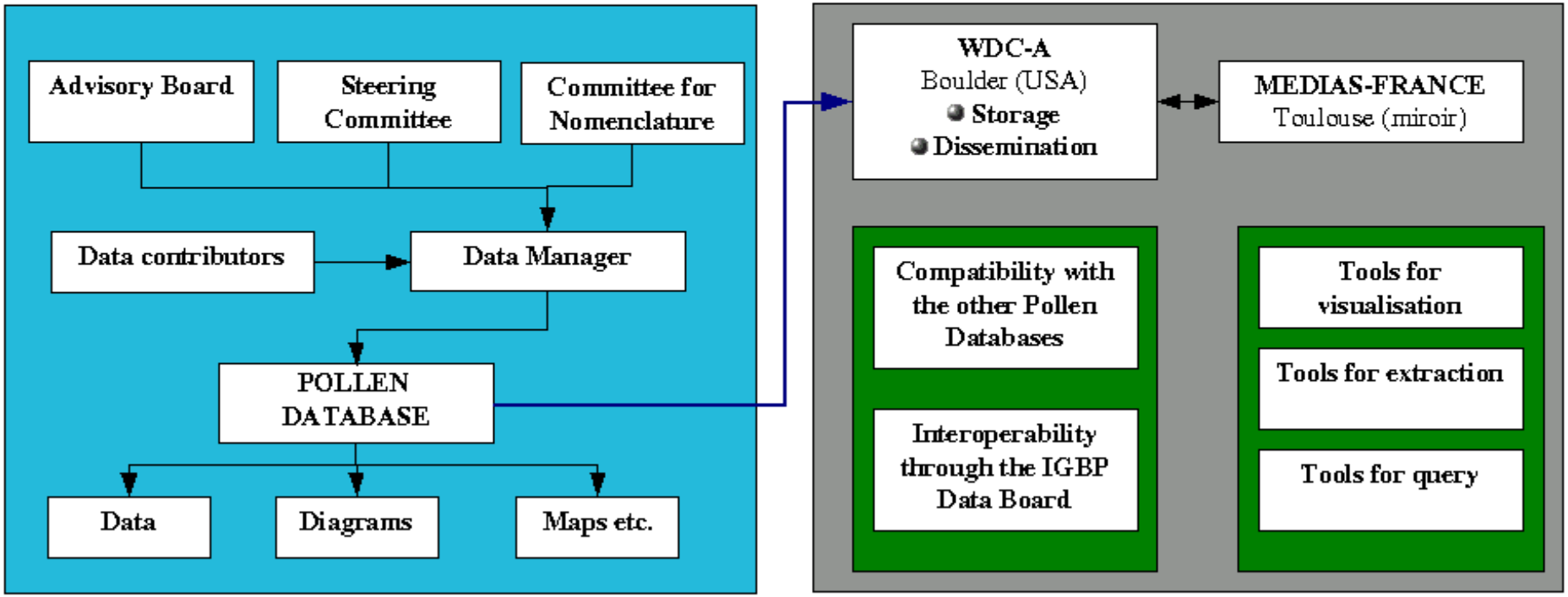
- To distinguish:
  - Scientific databases (experiments, networks)
  - Archives (Meteorological Services, BRGM, IGN, etc.
  - Collections (Museums, …)
- To question on the role of the different actors in the scientific world :
  - **the institutions of which the single goal is the scientific research** (CNRS-Insue, IRD, etc.),
  - **those of which the scientific goal coexists with the operational or commercial goal** (e.g. BRGM, IGN, etc.)

Organization of the community

- **In scientific centers**: network developments between scientists of different projects in order to adopt:
  - Same formats for common data
  - A common documentation (metadata) by discipline
  - Quality control informations
- **In technical centers providing**:
  - Database development and maintenance
  - Data distribution

Some examples of database organization

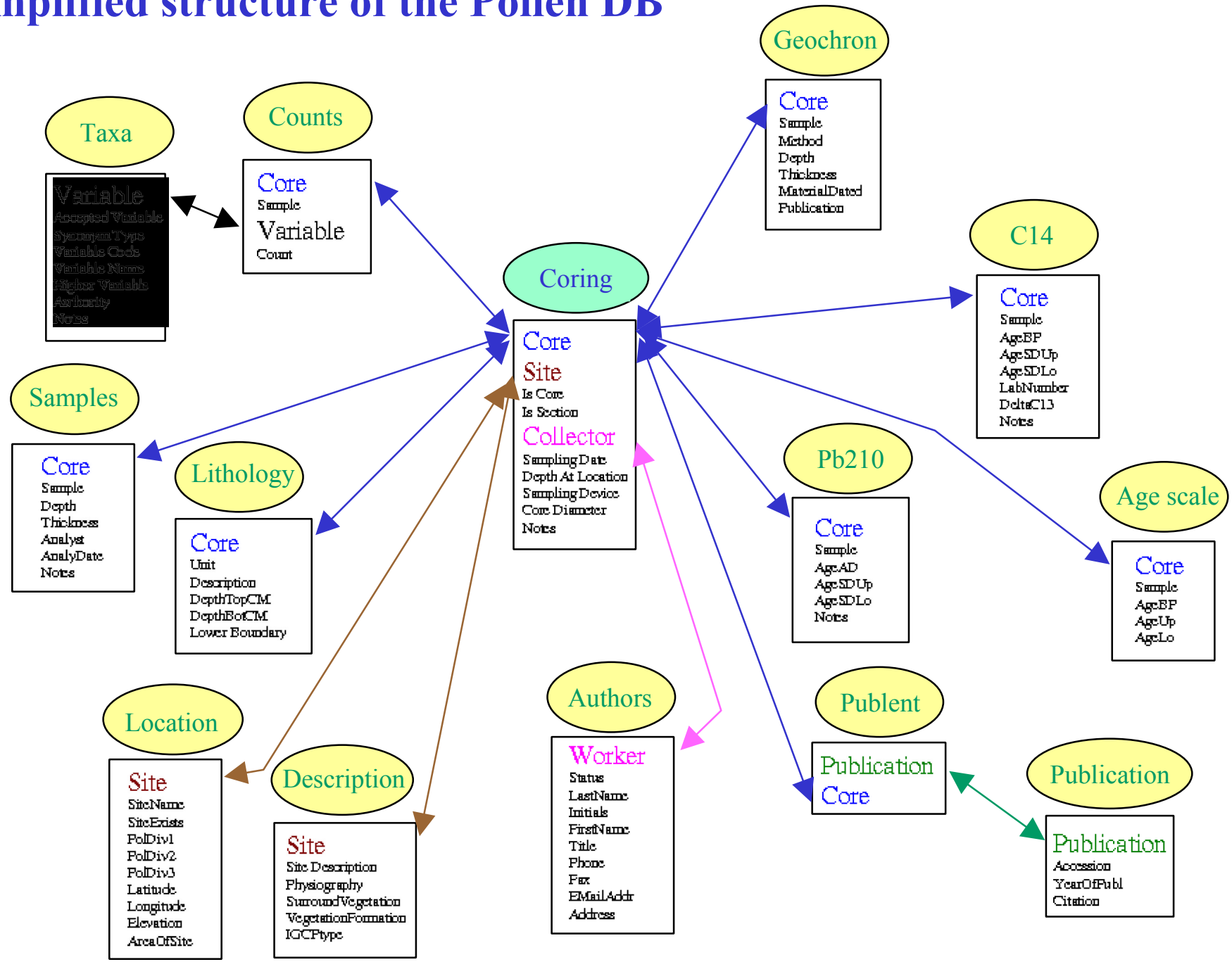
Scientific discipline database
An example with pollen data



Scientific entity

Technical entity

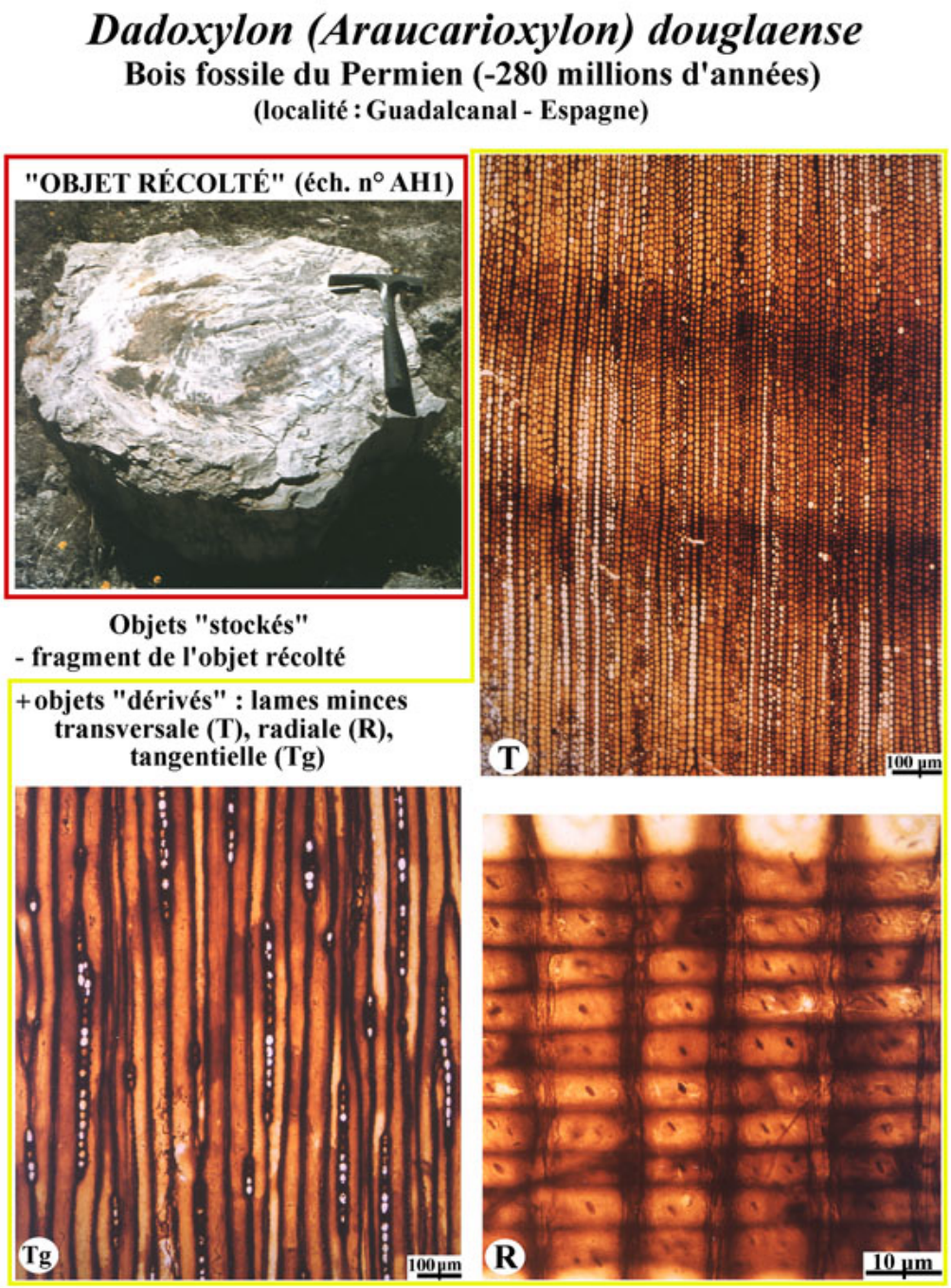
Simplified structure of the Pollen DB



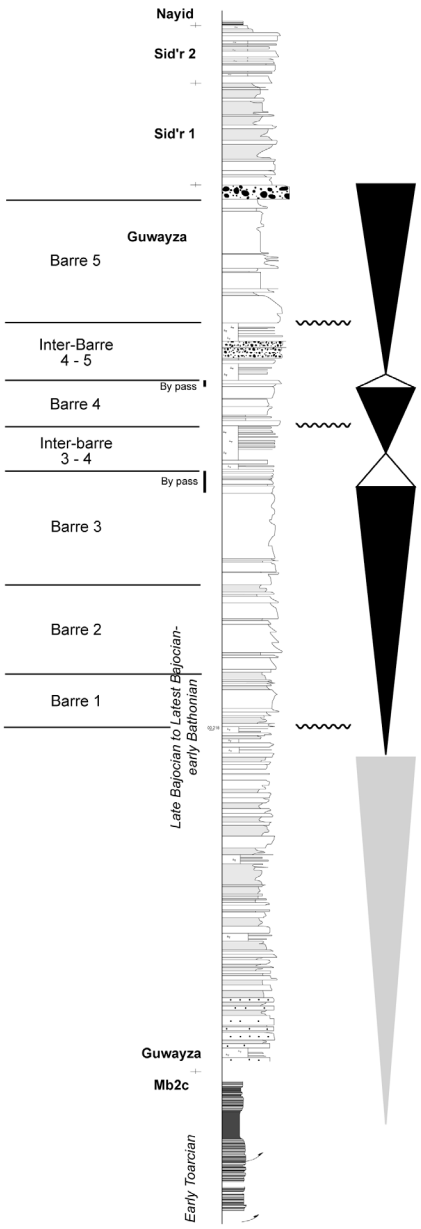
Examples of various data

Fossil wood

The taxa identification and the observation of tree rings give information on climate changes



The sedimentary faces



Guwaysah formation, Oman (F. Guillocheau, C. Robin)

Megarid 3D of gravity system

Conglomerate

PALBOT

- Base de données de la collection de Paléobotanique de l'Université Paris 6
- Environ 15000 fossiles végétaux de natures très diverses :
  - macro et micro fossiles ;
  - structures perminéralisées ;

empreintes et compressions

- Importance pour la recherche ; nombreux holotypes
- Base de données en cours ; accessible à l'URL *http://albinoni.snv.jussieu.fr*

Classification Evolution et Biosystématique
Laboratoire de Paléobotanique et Paléoécologie (J. Broutin)
Laboratoire Informatique et Systématique (R. Vignes Lebbe)

UNIVERSITE PIERRE & MARIE CURIE
LA SCIENCE A PARIS

Catalogues – Meta-données

• **Campagnes** : 5445 résumés **ROSCOP/CSR** (Cruise Summary Report)

• **Bases/Jeux de données** des laboratoires de la communauté françaises ou acquis à titre d'échange : 306 fiches descriptives **EDMED** (European Directory of Marine Environmental Datasets) (dont 82 au SISMER)

• Stations d'observations « Temps Réel » **EDIOS, MAMA**

• **EUROSEISMIC**

Serveur FETCH/CETP - Netscape

File Edit View Go Communicator Help

Back Forward Reload Home Search Netscape Print Security Shop Stop

Bookmarks Location: http://dataserv.cetp.ipsl.fr/FETCH/ What's Related

Internet Nouveautés Avoir Membres Connexions Marché

Documentation on the experiment
- General informations
- Means of observation
- Models
- Satellites
- Meteorological network

The experiment
- Overview of measurements
- Progress reports
- Daily reports
- Annex

Fetch database
- Fetch data policy
- Data described by DIFS
- Access to the data

Reports, Publications
- Internal Reports
- Publications

Miscellaneous informations
- Server user's guide
- Photographies
- Acronyms

Last updated: 23/11/01

This server is developed and maintained by CETP

Welcome to the Fetch home page

This is the server relative to the Fetch experiment

Flux, Etat de la mer et Télédétection en Condition de fetcH variable

MEDIAS

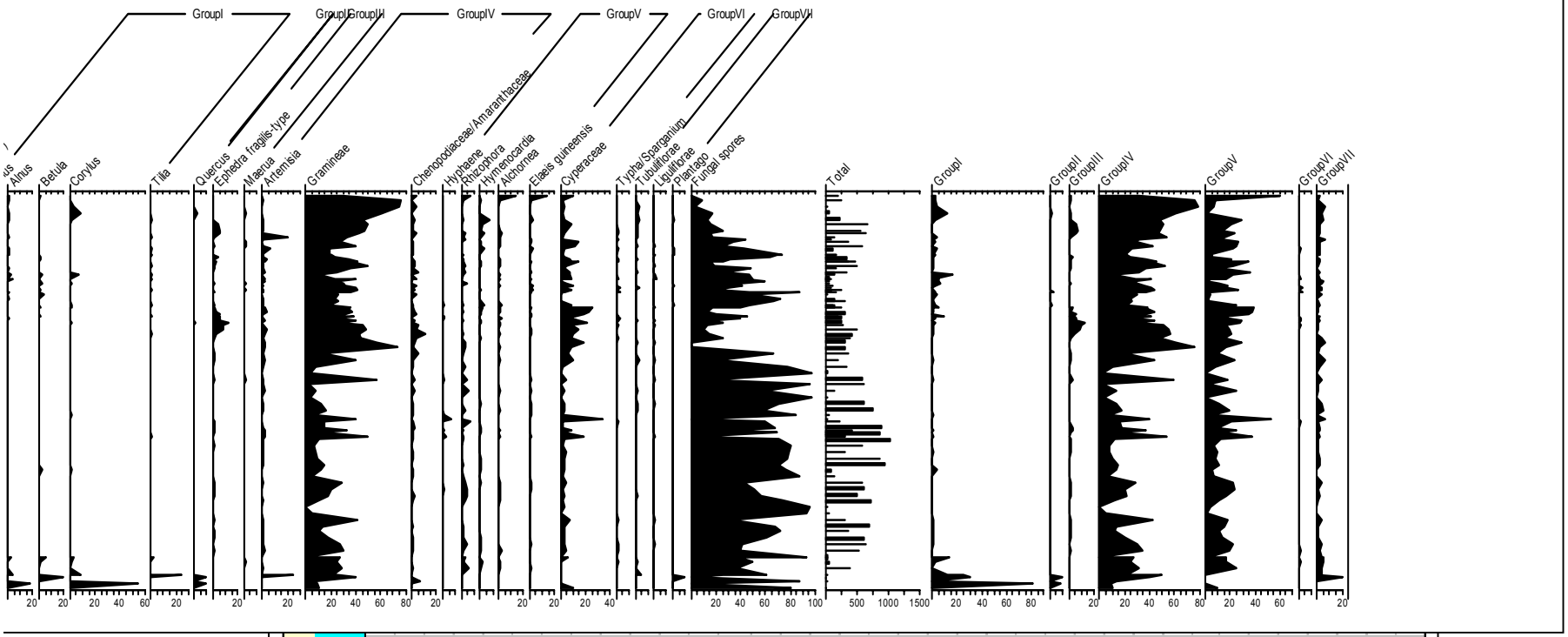
INSU METEO FRANCE Ifremer CNRS

MAST - III
Marine Science and Technology Programme

Document: Done

Démarrer | Microsoft Word... | Microsoft Excel... | Explorateur - C:... | Eudora Light - [... | Les Bases de d... | Microsoft Powe... | INCO-APD - Ne... | Serveur FE... | 12:13

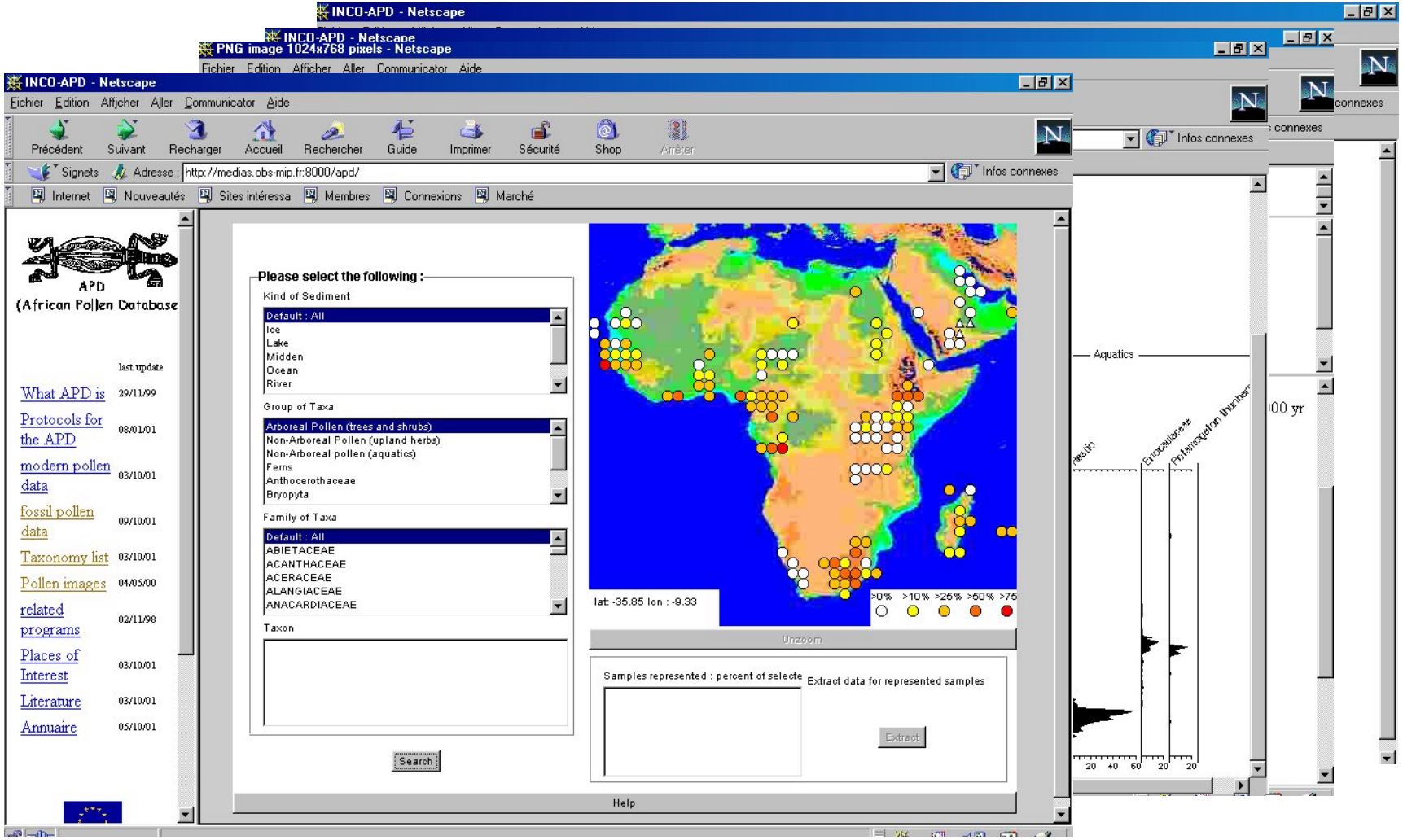
Les diagrammes polliniques



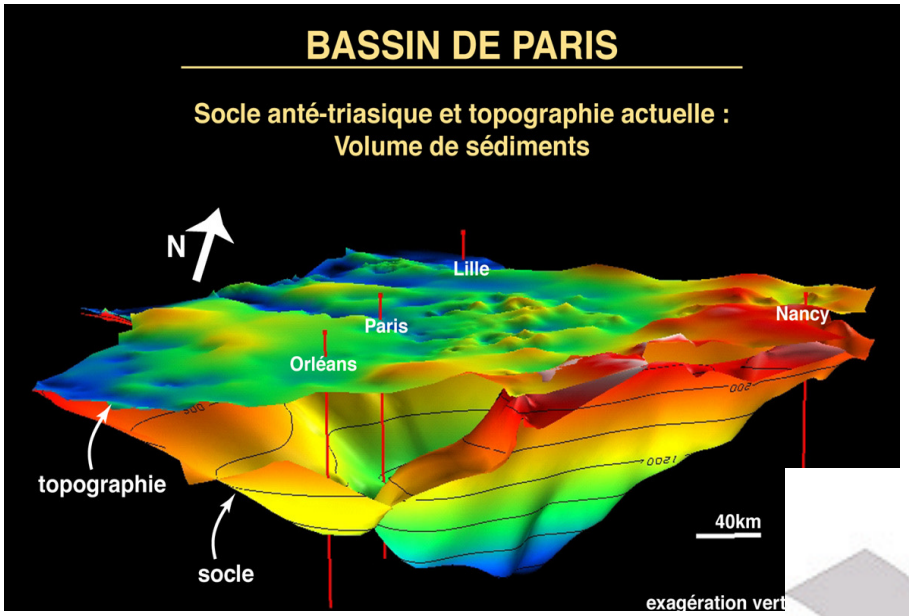
674	11959	0	0	3	0	1	0	1	0	1	2	0	0	132	0	0	0	1	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0
700	12311	0	0	6	0	0	0	0	0	0	0	0	0	149	0	0	0	0	0	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0

Juillet-2001 Jean-Pierre Cazet

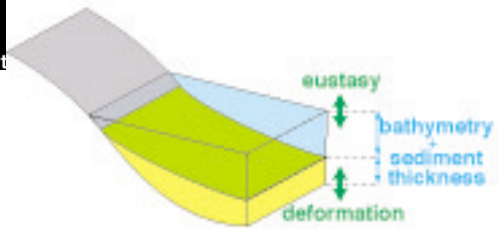
The African Pollen Database



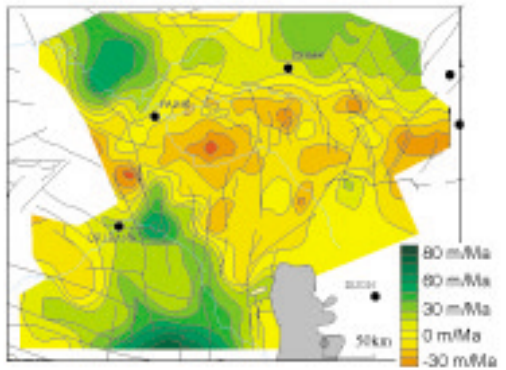
The numerical simulations



3D Visualization



Quantification

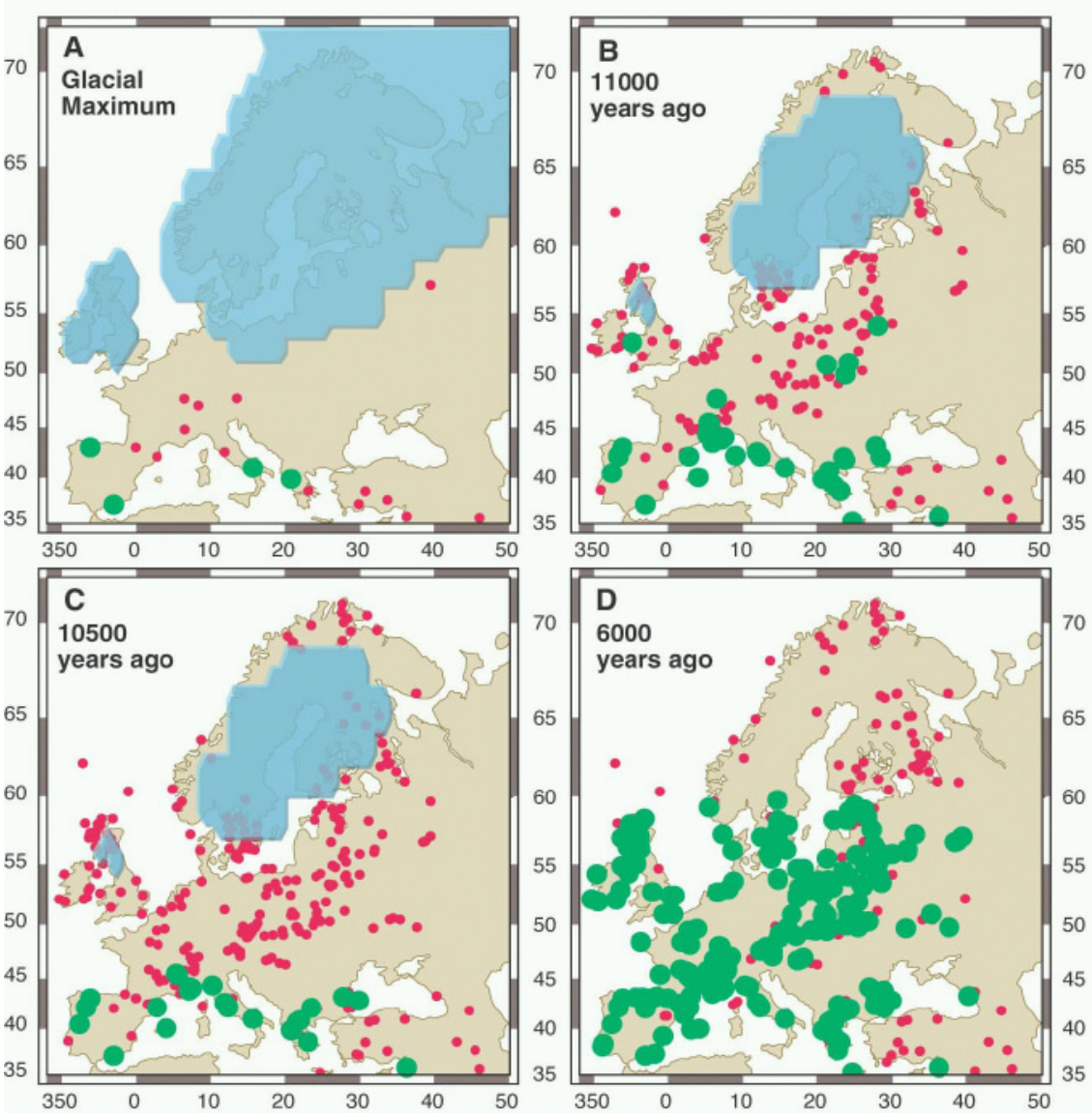


F. Guillocheau, C. Robin, D. Rouby

Data valorization

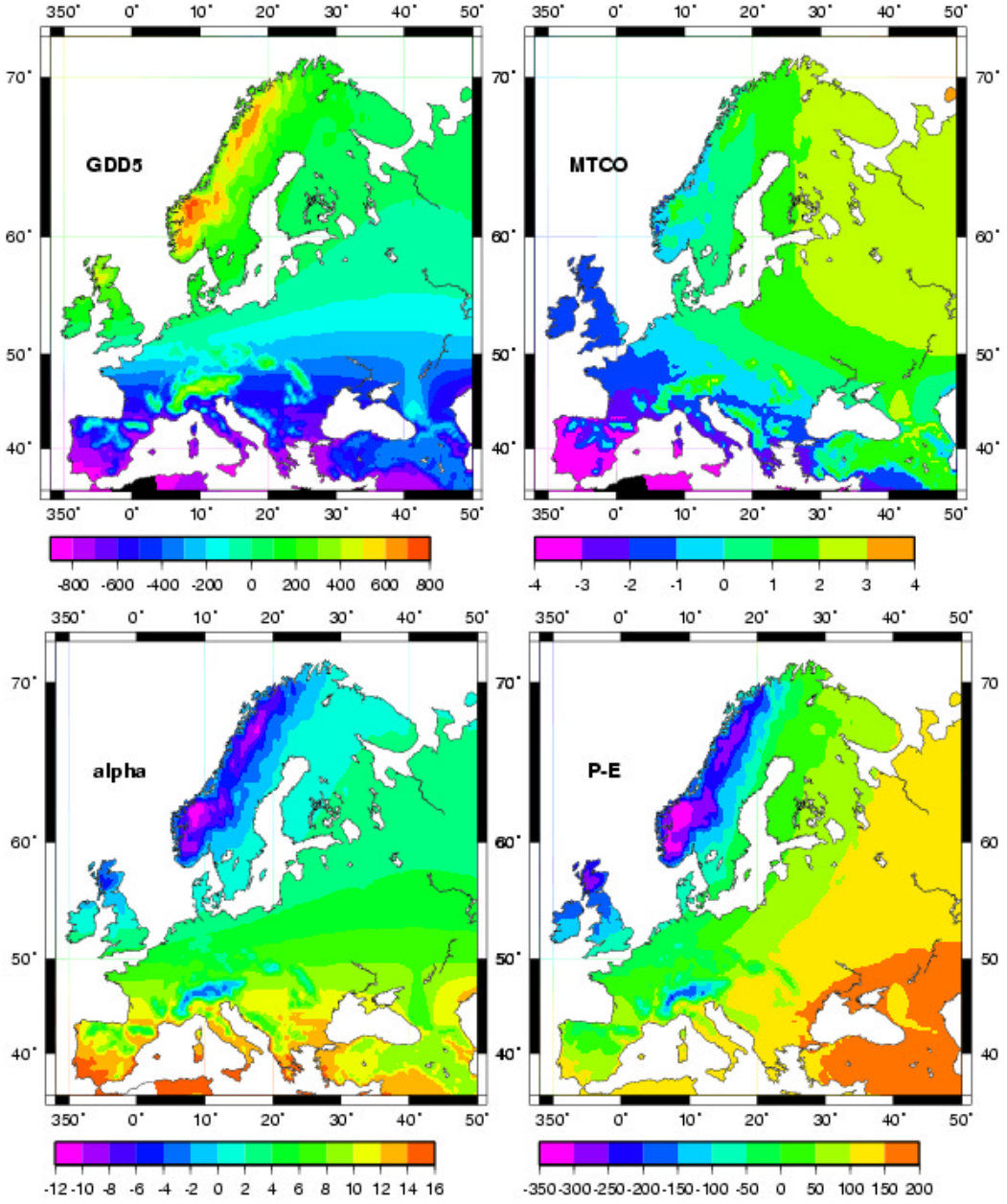
European Pollen Database

Oak migration in Europe during the last deglaciation



Taberlet et Cheddadi, *Science* (2002)

Climate reconstitution 6000 years BP



(Cheddadi et al., *Climate Dynamics,* 1997)

Objectives given to the databases

- **Data archiving, which implies a discussion on the stakes and the duration of it.**
- **Data access**, which the services to be given to the users, with the realization:
  - of a data policy by the scientists
  - by the technical operator, in strong relationship with the scientists:
    - of a data visualisation and extraction interface
    - of data valorization tools

A coordination to be obtained between the institutions

- Database functioning rules
- Data value and **property**
- Relationship with industry for an open access to the data
- Setting up of a data policy

Planned

- The setting-up of a **data portal**:
  - This implies the adoption of international standards for metadata formats
  - This allows a flexible organization for data management

- The starting up of a community **(inter-labs and inter-institutions) working in database interoperability**

International standards for the metadata

- In order to develop a portal for a common access to the different databases, the priority is to collect all the **descriptive information** on the data which will allow to use it.
- **Numerous French institutions** are working on, particularly Cnes (space), Medias-France, BRGM (geology), Ifremer (oceanography), etc.
- The international metadata used by them is: **"ISO/DIS 19115 Geographic information -- Metadata"**

A federated and adaptable data management

In this view, shared by Medias-France for the development of « multi-proxy » databases for scientific projects:

- **The data expertise** is staying in the laboratory which has produced it
- **The data center centralizes the different initiatives and manages the coordinating system based on a metadata catalog**
- **The user** asks one server for the access to the data coming from different databases initiatives

Objectives

- To make easier the efforts to archive the scientific data
- To allow to describe and to localize them
- To homogenize the ways of access and use

Data archiving

- Data are saved in centers specialized in the data storage
- Data are coherent and easily updated
- Access policies are under the control of the scientists in the data centers

Facilities for the search and the localisation of the data

- Data are described by the metadata in a standard format
- Tools used for the description of the data are easy to use by the data providers (Web forms)
- Management and data access policy of the metadata are provided by a single entity

Access facilities

- Data pre-visualization and extraction are standardized
- Data is not redundant
- Data access policy is managed by the data providers

Organization (federating unit)

- A center has to manage the metadata
- It gives data centers the necessary tools for data description
- It provides the scientific community the tools for data search, description and localization

Organization
(Scientific Data Centers)

- The data centers manage the data with their methodologies
- They implement (with the help of the federating unit) a standard interface for data exchange
- They describe their data with the standard tools provided by the federating unit.

Script of data exploration and extraction

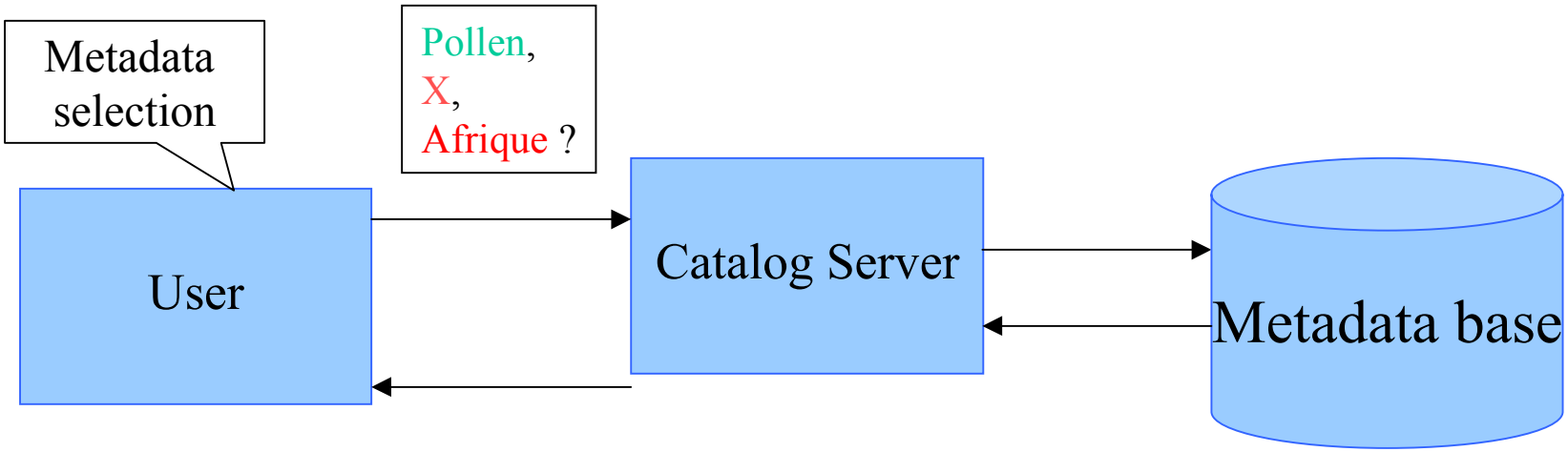
The metadata

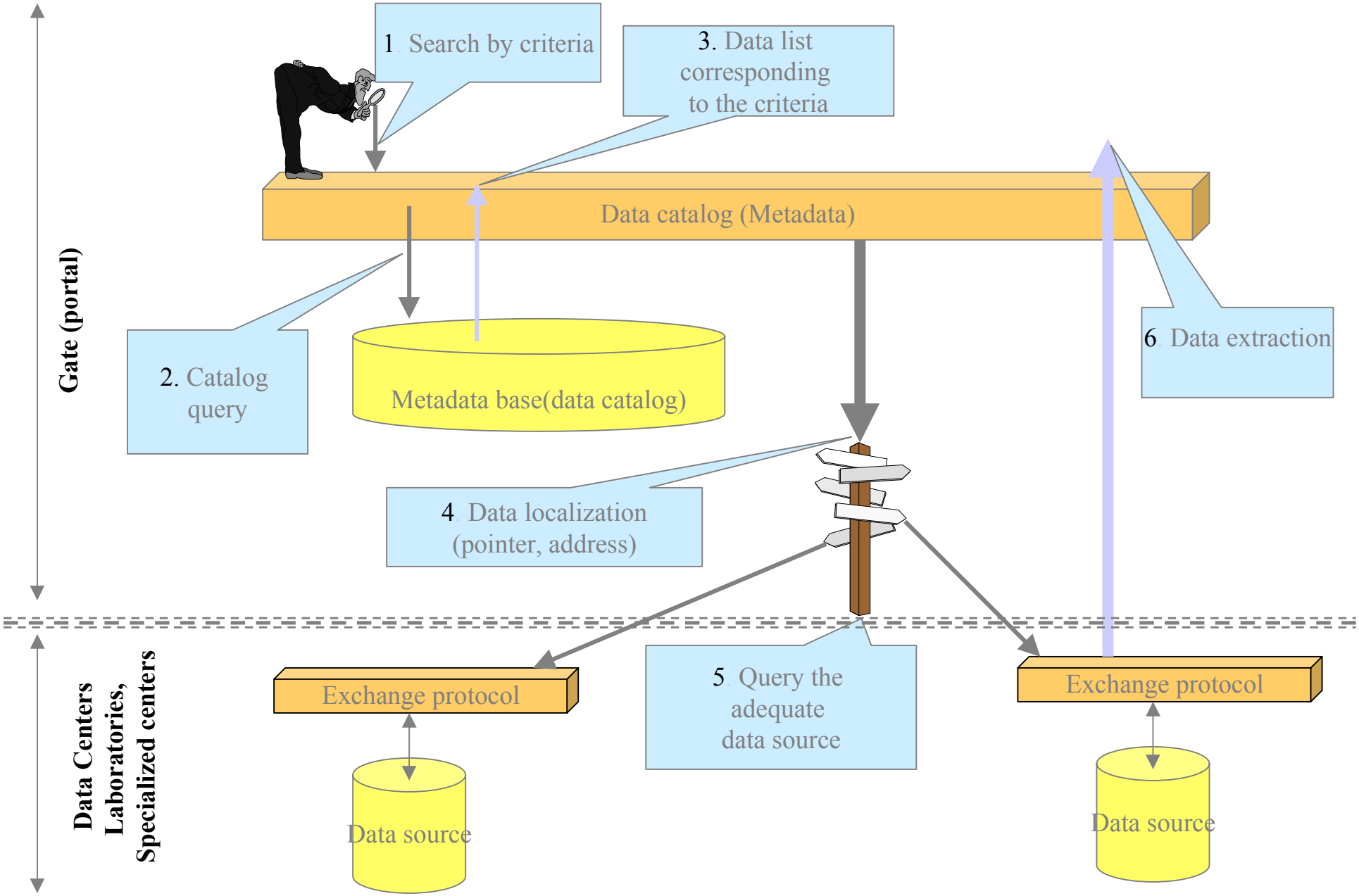
- *Problem*: the user would like to access to the data via their metadata

- *Example* : « I would like to know the pollens sudied by M. X in Africa »

One solution: the catalog server

- Principle :




Gate (portal)
1 Search by criteria
3. Data list corresponding to the criteria
Data catalog (Metadata)
2. Catalog query
Metadata base(data catalog)
6 Data extraction
4 Data localization (pointer, address)
5 Query the adequate data source
Exchange protocol
Exchange protocol
Data source
Data source
Data Centers
Laboratories,
Specialized centers

Medias-France proposal

- To create ISO 19115 profiles for various scientific disciplines, in strong relationship with:
  - The Pi's of scientific disciplinary programs (GPD, etc.)
  - The responsible of scientific international and interdisciplinary programs (WCRP, IGBP, IHDP, etc.)
- To install a catalog server which will exploit these profiles

The data servers

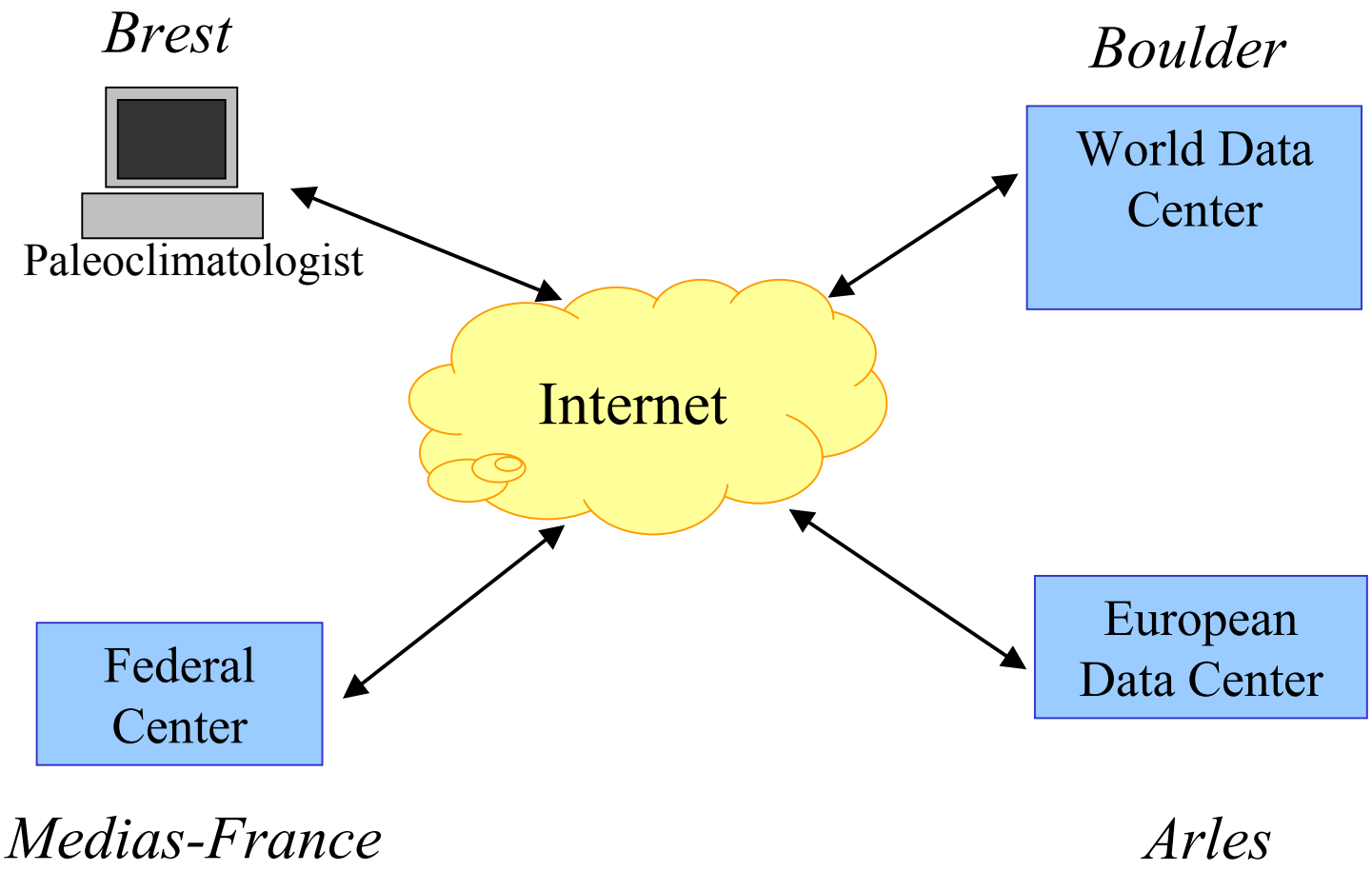
- Are accessibles via the catalog server
- Allow data visualization
- Example : the x-proxy server for the paleoclimatology

The x-proxy server

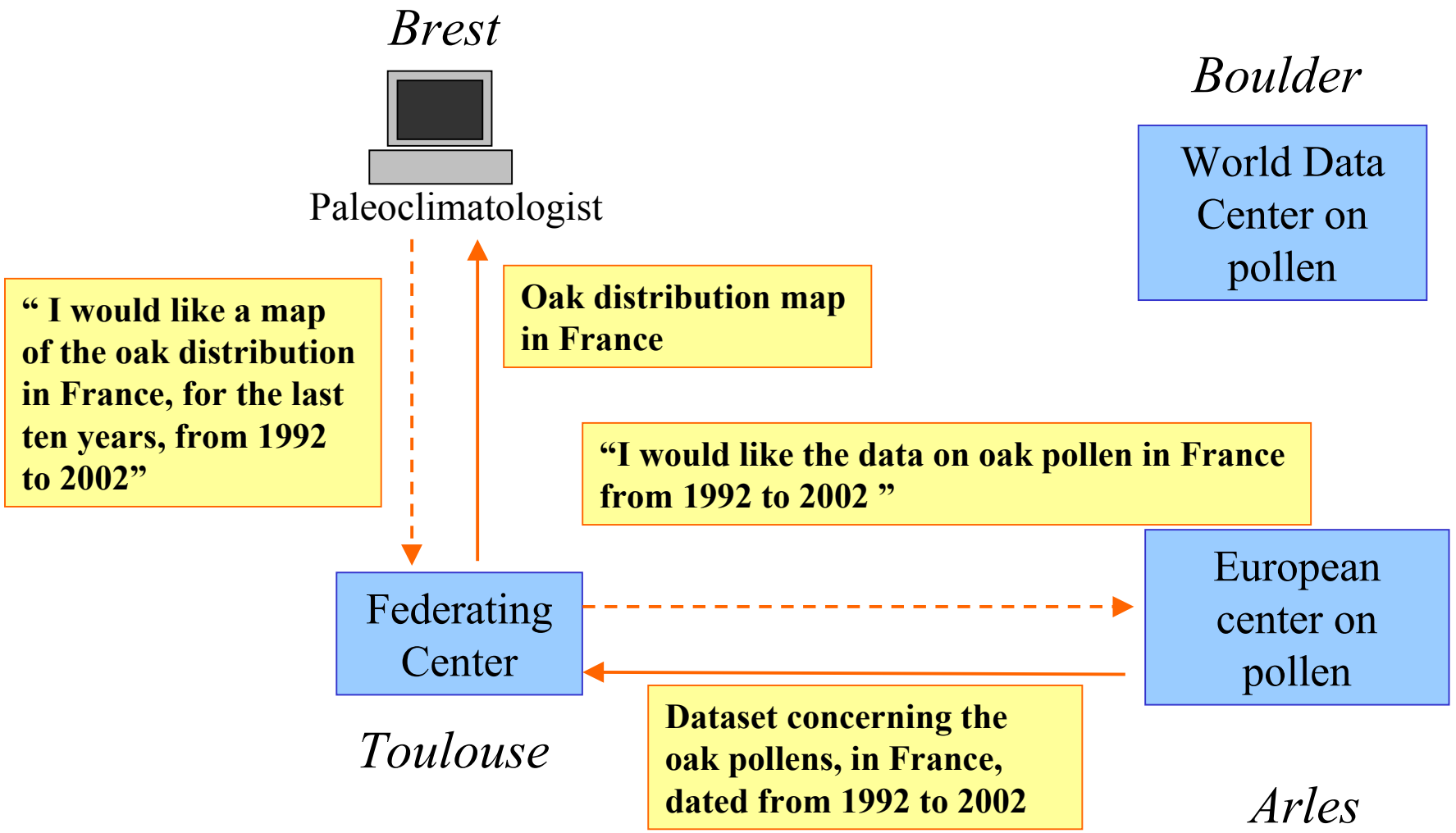
- The need: to reach with a single interface heterogeneous and remote data

- The condition: data ownership and management is done by the scientists

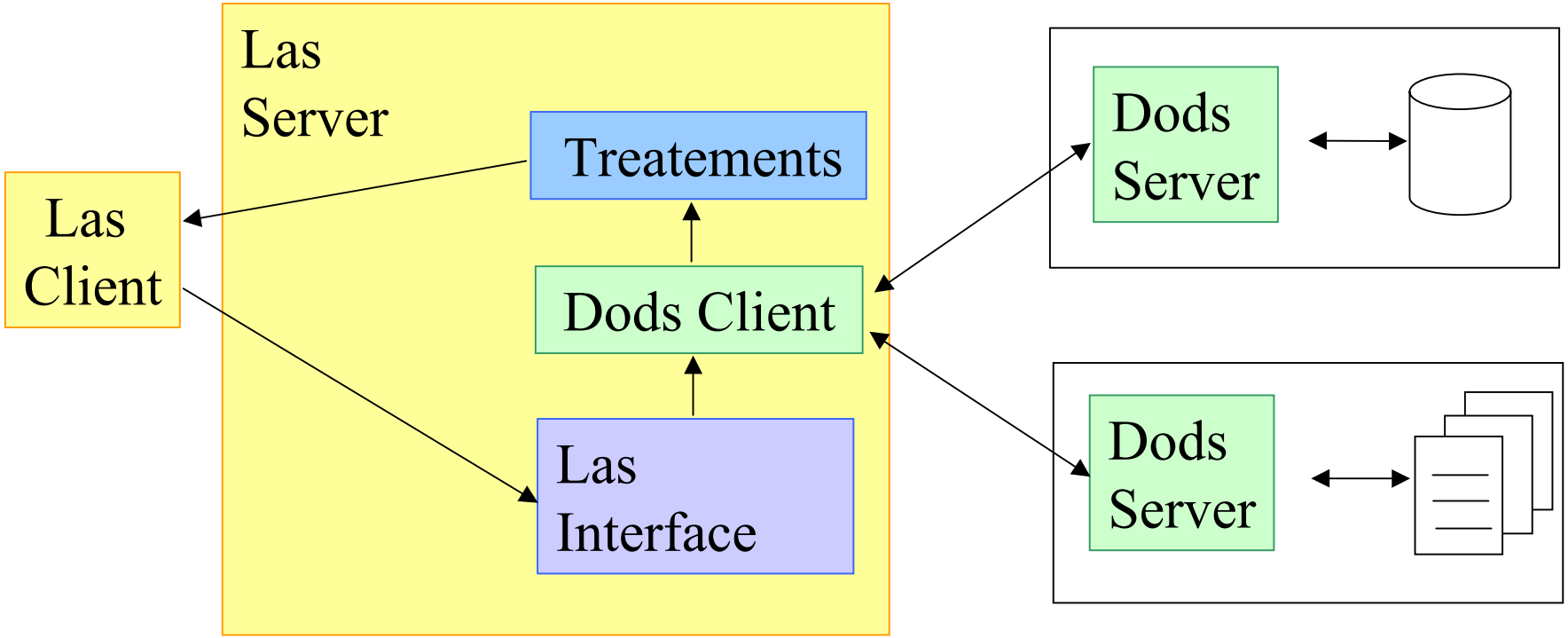
An example of an answer to this need



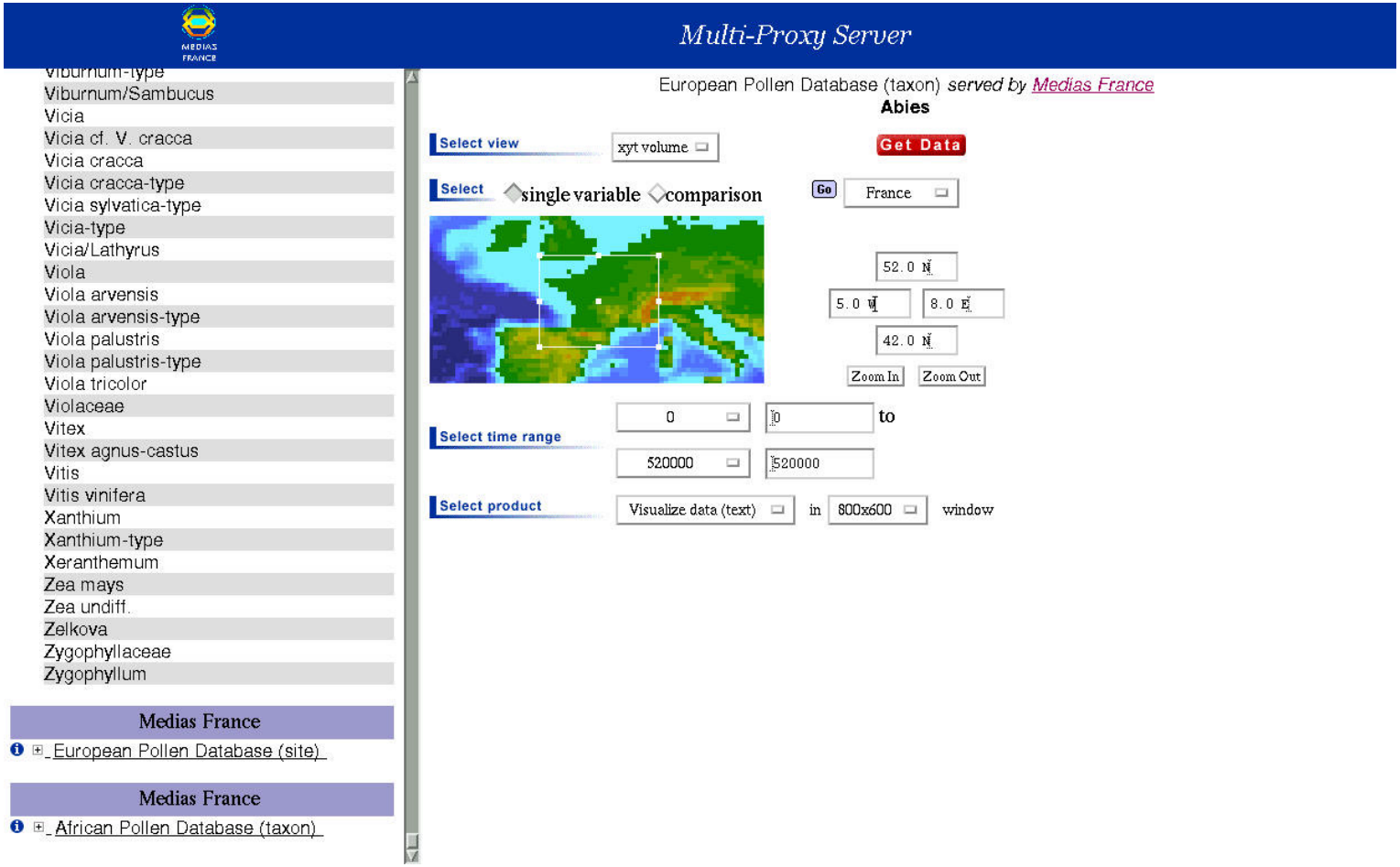
Exchanges example



Las-Dods Architecture



An example of « multi-proxy » database



Proposition

- **To archive** the pertinent and essential data, validated by the PI

- **To manage and maintain** the databases in technical agencies like Medias-France and in research teams

Specific role of Medias-France

- To assure that the data will be stored after their validation by the scientific community
- To propose the specific access and management tools for each type of data
- To develop the "multi-proxy" interface

References (and acknowledgments)

- Communication of Anne-Marie Lezine at the Prospective Symposium of the Insu Division « Earth Sciences » the 24th of September 2002 at Vulcania

- Waldteufel's report: «Les bases de données pour les Géosciences, éléments d'un schéma directeur » published by the INSU et the CNES in 1999 (http://medias.obs-mip.fr/www/francais/documentation/)

Medias:

A votre service