Best practices for data sharing: metadata, data policies and databases.

Laurence Fleury, SEDOO, OMP, Toulouse

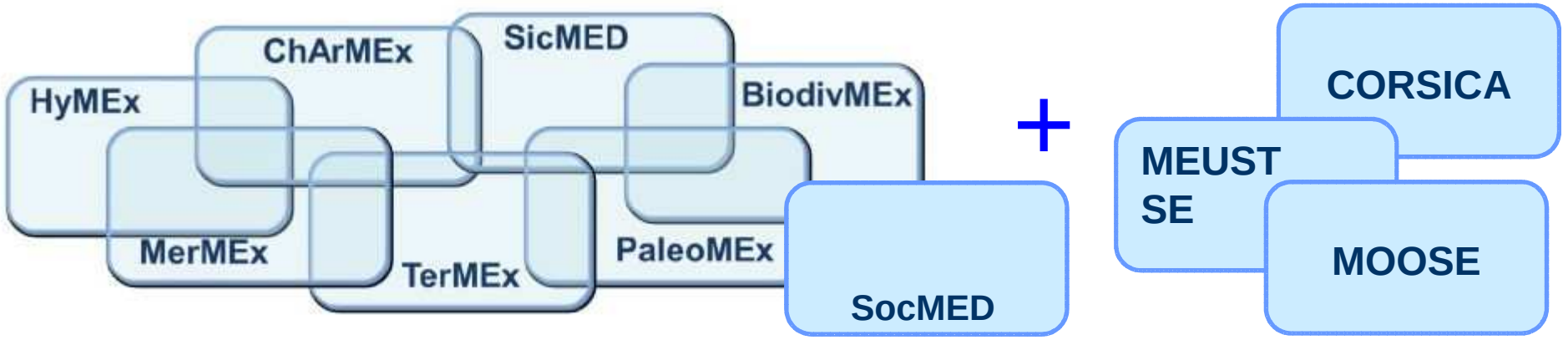
Two illustrations:

- The **MISTRALS** data portal **http://mistrals.sedoo.fr**
- The **RBV** metadata portal **http://portailrbv.sedoo.fr**

Data management basics

- **Metadata / catalogue** (homogenized data description)
  - who, why, where, when, what, which instrument, use constraints…
  - Standards: ISO, OGC, INSPIRE directive, thesaurus…
  - DOI (Digital Object Identifier) referencing
- **Data homogenization**
  - Choice: original data files *vs* set up of a relational database
- **User friendly data access interface**
  - **Interoperability**: a single portal to access distributed databases,
  - Easy search, select and download online tools
- **Data exchange rules, data policy / User registration**
- **Administration tools**
  - User registration request evaluation
  - Data access monitoring
- *Optional extra:* **Advanced tools** (visualisation, calculation…)
- **Back up**, long term storage

data portal context



- **Many scientific thematics and communities**
  - Different data sharing cultures, intellectual property definitions
  - Different methods, tools, standard formats, properties names and units, quality flags, user expectations…
- **Data heterogeneity**
  - *In situ* observations, satellite products, model outputs, social sciences surveys…
  - Research data *vs* operational data
  - Many time scales: historical data, observatories, campaigns…
- **Several data centers involved**

Principles

- *Challenge*

  A cross-disciplinary Information System involving all the MISTRALS projects.

- *Objectives*
  - Scientific data and information exchange
  - Knowledge diffusion to a general public (decision-maker dedicated products, vulgarization pages, GIS tools…)

- *Tools / present status*
  - Catalogue, data dissemination / operational tools
  - GIS / to be defined and set up

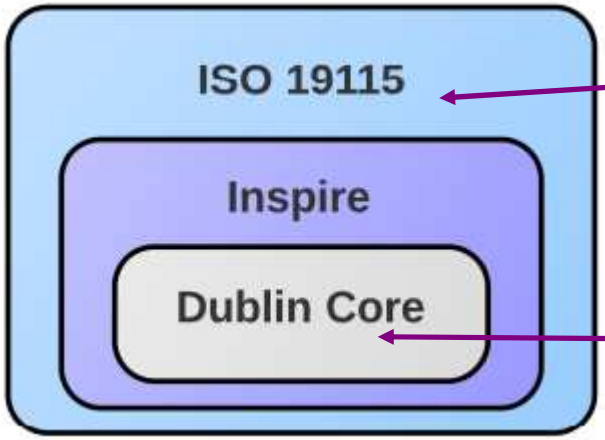
The **RBV** metadata portal

- RBV - Réseau des Bassins Versants- is a network of 15 observatories concerning watersheds monitoring.

- Needs:
  - a metadata portal to document the observatories and the datasets.
  - a tool to help the different groups to communicate, collaborate and homogenize data acquisition procedures.

Metadata

- **ISO norms** and **INSPIRE** European directive:



ISO 19 115, 19 115-2
About « geographic information ».*
Unlimited information, different profiles.
ISO 19139: xml implementation scheme

15 fields: title, subject, creator…

- **Open Geospatial Consortium** standards about geographic information
  - Interoperability protocols: WMS, WCS, WPS, CS-W
  - Description standards: O&M (Observations and Measurements),SensorML…
- ISO vs OGC: complementary, **but** sometimes conflicting…
- **(multi)disciplinary thesaurus** to reference datasets.

MISTRALS metadata module

- Catalogue based on a **relational metadata database**
  - designed together with other data centers (SEDOO, ICARE, IPSL)
  - including **all the datasets** (*in situ*, gridded product…)
  - complying with **ISO19115 norm**, **OGC standards**, **INSPIRE directive…**
  - parameter and instrument keywords from the **GCMD thesaurus** (Global Change Master Directory).
  - thesaurus to be extended in case of socio-economics, demography… datasets.

- **Metadata form** to be filled by the data provider or data centre.

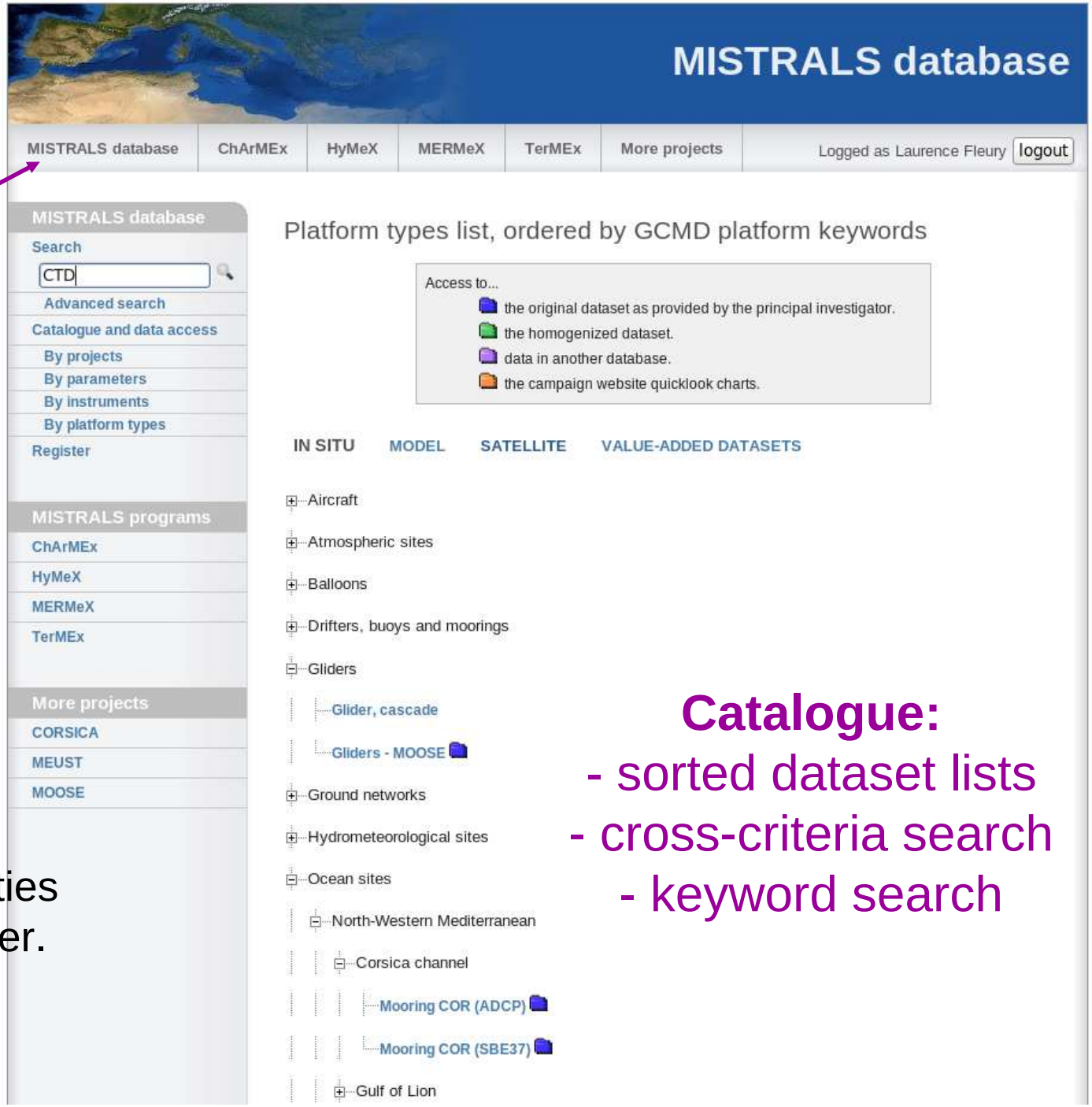
- **Metadata edition:**
  - Every filled information
  - Measurement location plotted in a map
  - Data access
  - DOI (only if data are available).

http://mistrals.sedoo.fr/

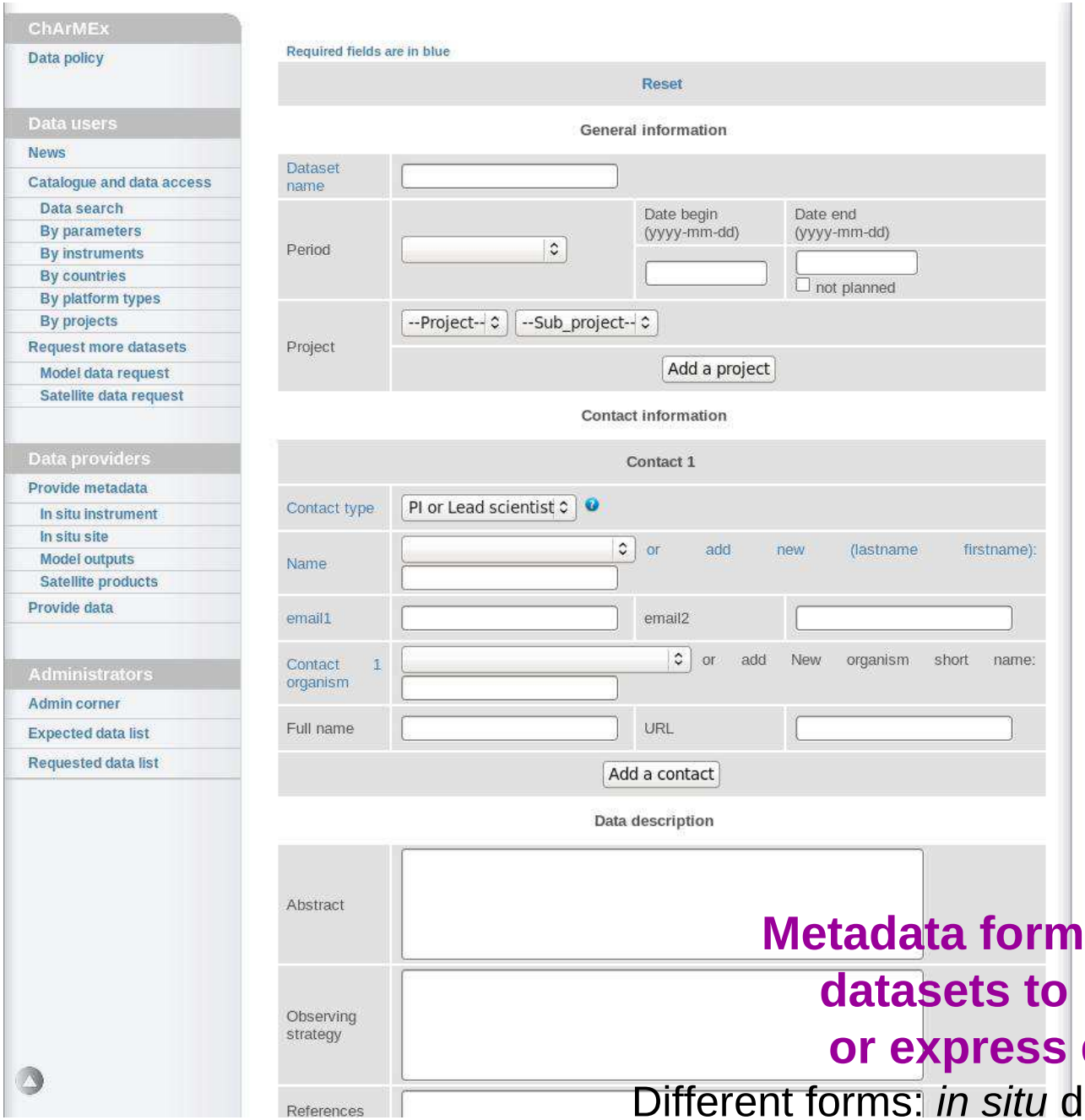
Every data

One tab
by project

Similar fonctionnalities
but menu may differ.



Catalogue:
- sorted dataset lists
- cross-criteria search
- keyword search



MISTRALS Metadata

Metadata forms to document datasets to be provided or express data needs.

Different forms: *in situ* data, site, model outputs, satellite products, value added products

Metadata edition

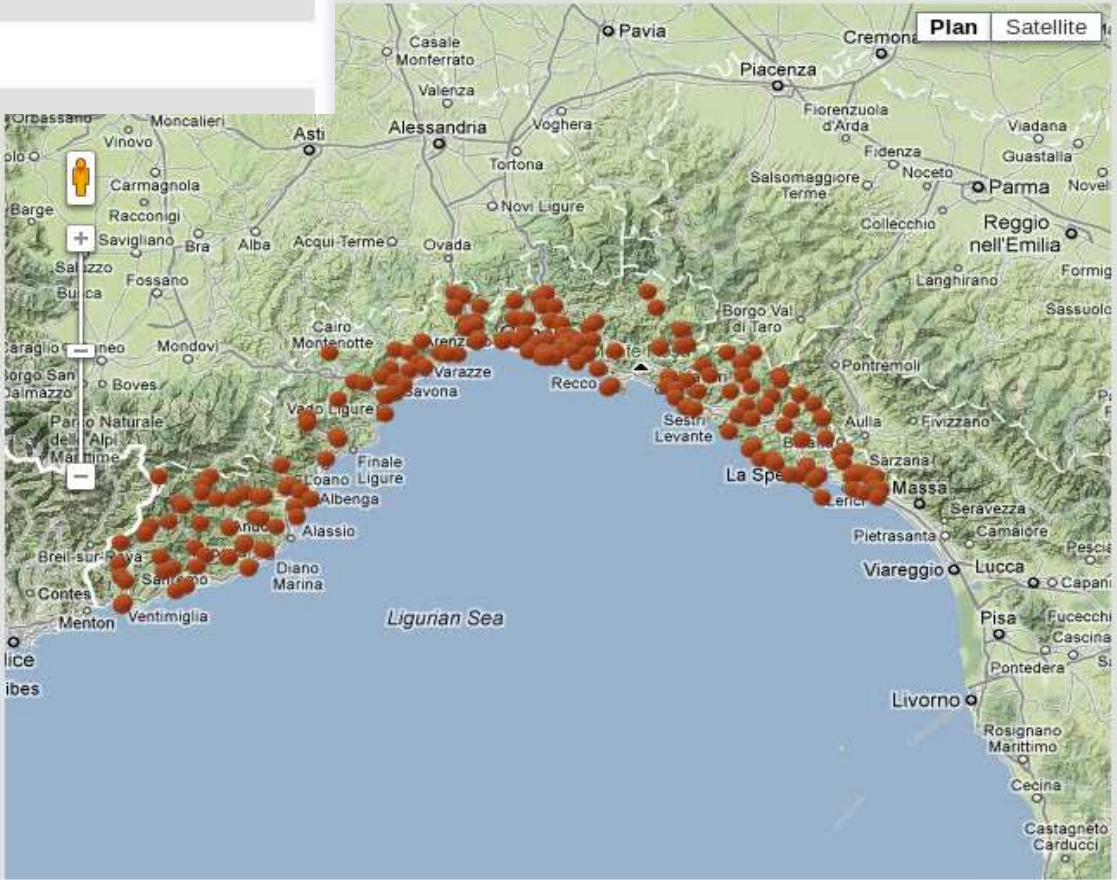
General information

Dataset name	Hourly surface stations, Liguria		
Created on	2012-05-03		
Project(s)	HyMeX		
Period	LOP		
Date begin (yyyy-mm-jj)	2011-09-19	**Date end (yyyy-mm-jj)**	2011-11-10
Contacts	Sacchetti Davide - ARPAL (Dataset contact)		
Data access	Original dataset as provided by the Principal Investigator Homogenized dataset		

Data description

Geographic information

Location 1
uscany site (LT)
eteorological sites

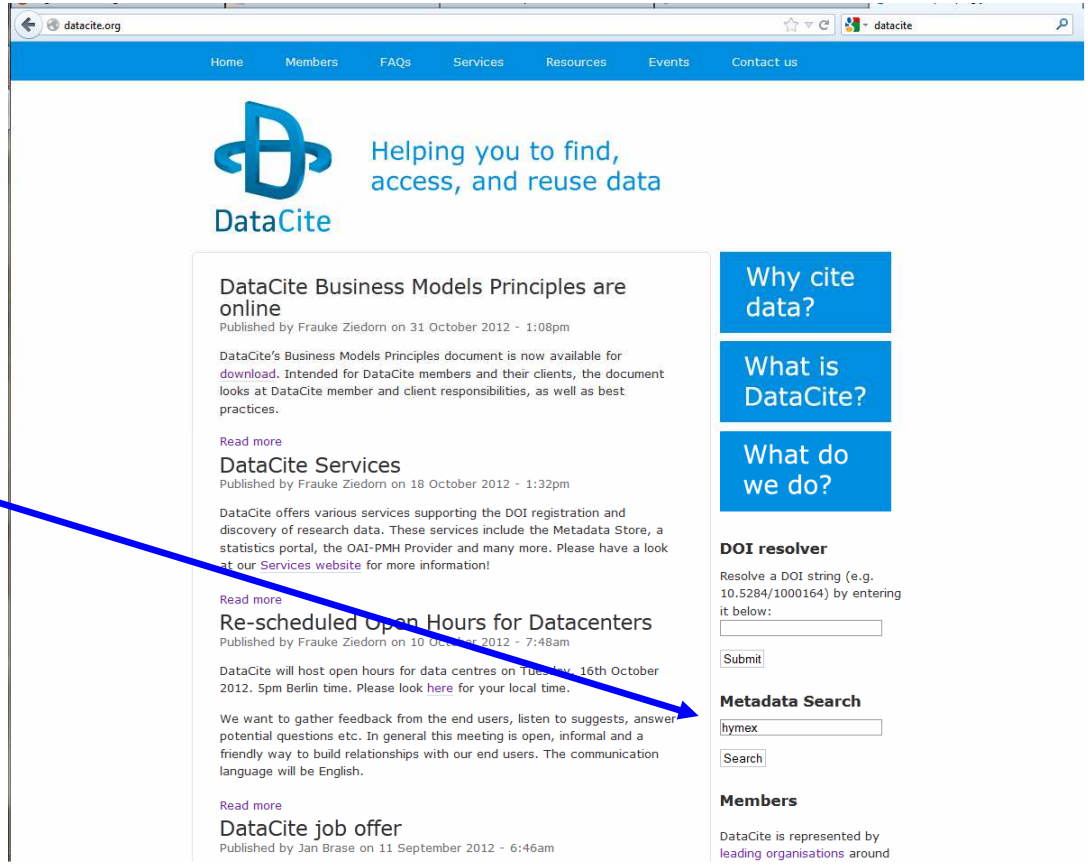


DOI (Digital Object Identifier)

- Dataset referencing allows an improved visibility for the scientist and dataset easy citation.
- Convention between DOI structure and data centre → commitment to data long term durability.
- Portal **DataCite**

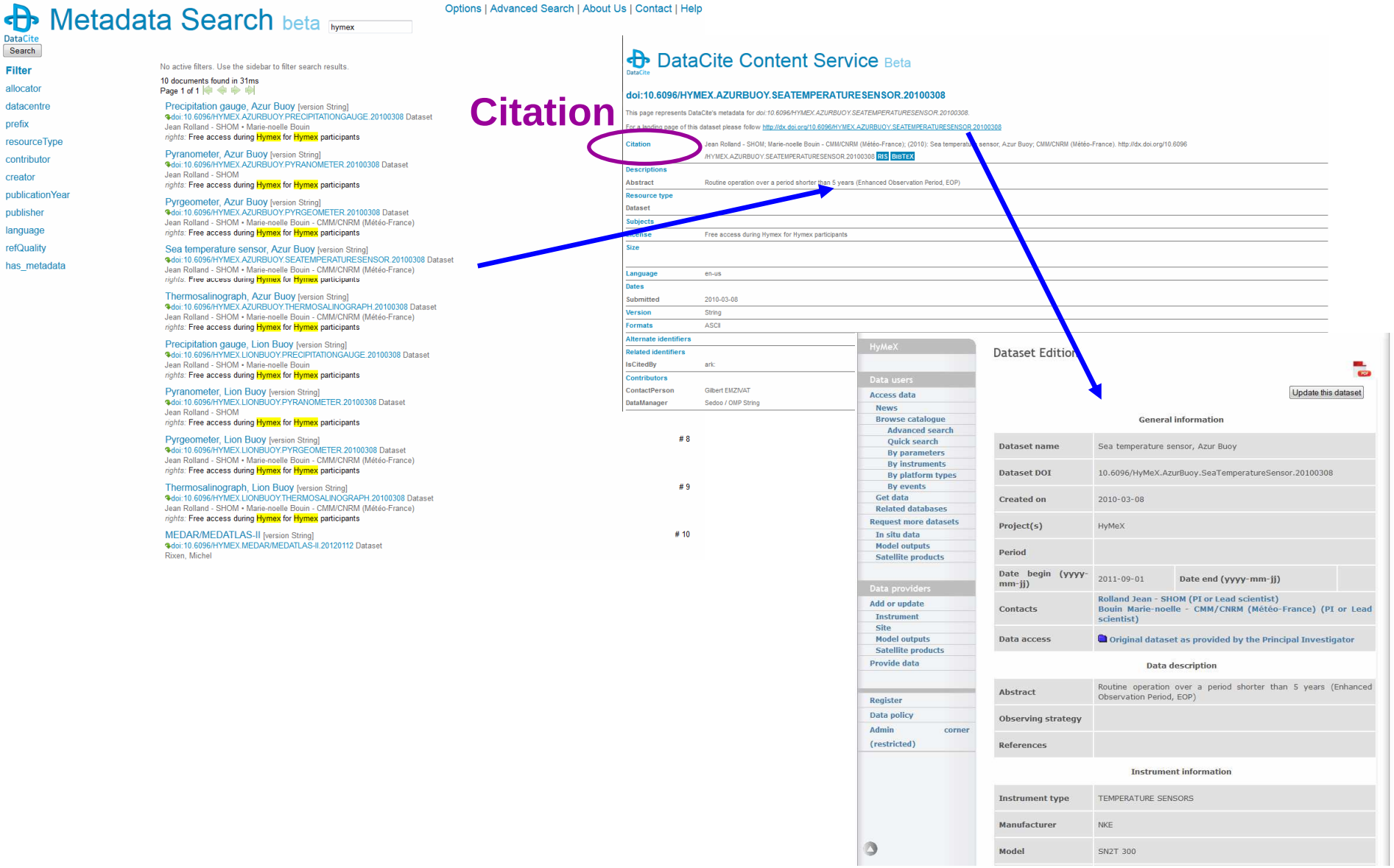
(all the resources referenced with a DOI except publications)

www.datacite.org



DOI referencing

doi:10.6096/HYMEX.AZURBUOY.SEATEMPERATURESENSOR.20100308

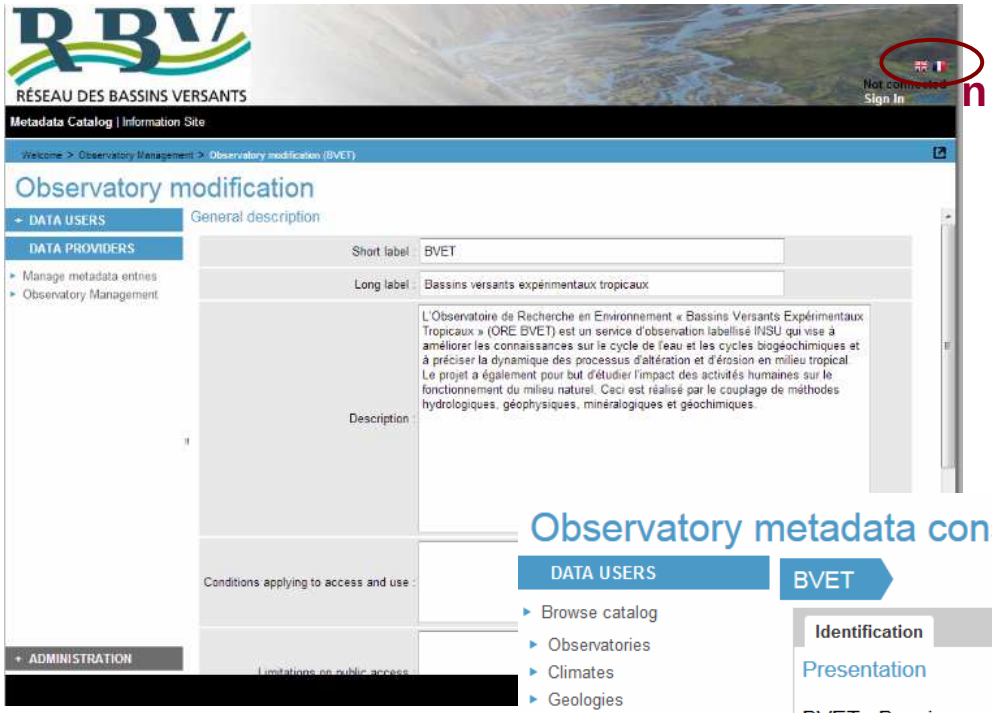


The **RBV** metadata portal

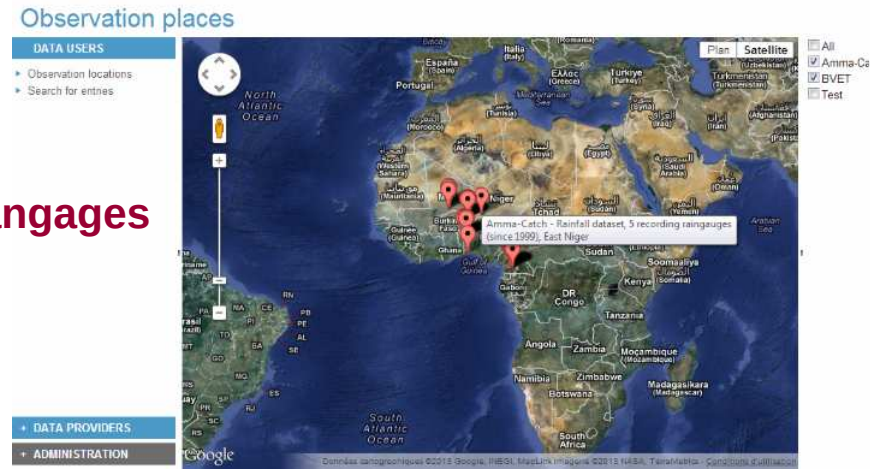
- **Three levels of metadata**:
  - Observatory
  - Experimental site
  - Dataset
- Use of a **catalogue software** (Geosource / Geonetwork) to store, search, harvest the metadata ; definition of a data profile compliant with ISO norms, INSPIRE European directive and OGC standards ; use of 52North for SensorML
- Developpement of a Java/GWT **specific online interface:**
  - metadata forms
  - metadata selection and edition
  - News, project information…
- Developement of a web service to generate ISO 19139 metadata from a simplified RBV format.

http://portailrbv.sedoo.fr

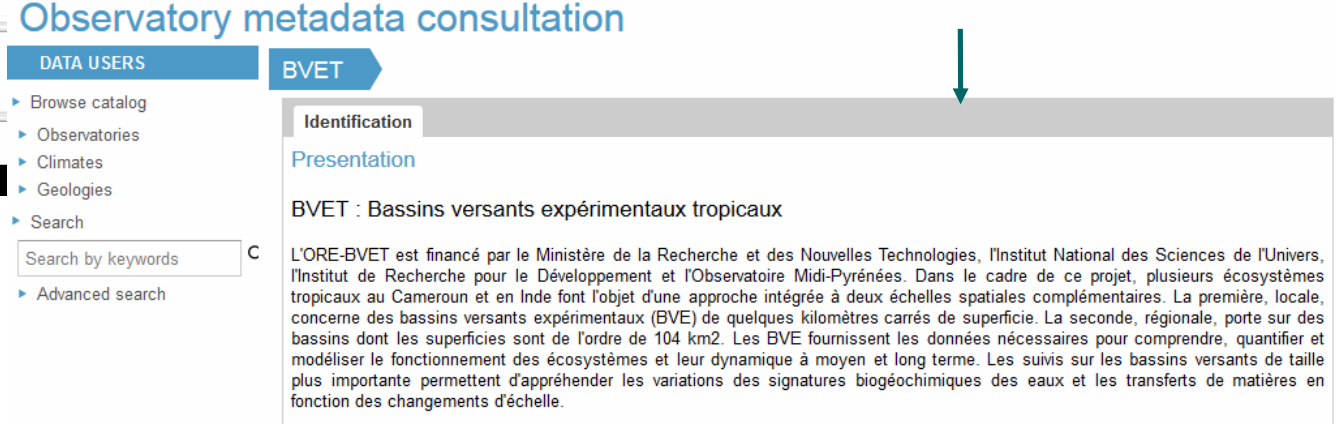
Form to fill observatory information



n langages



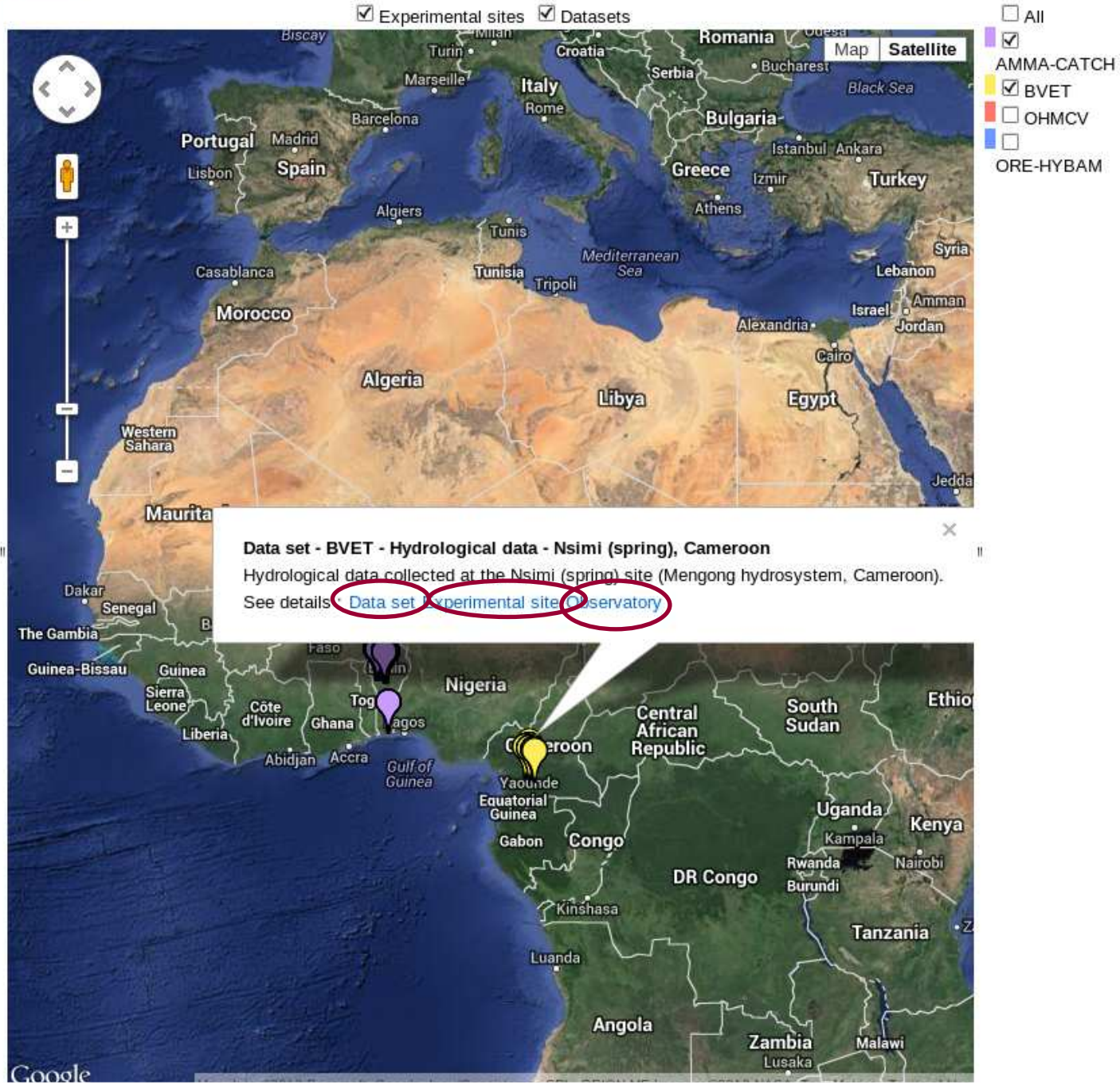
Observatory metadata consultation



Observation places

DATA USERS

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search
- Advanced search



Observatory metadata consultation

DATA USERS

BVET

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search
- Advanced search

Identification | Contacts | Keywords | Geographical Informations | Temporal informations | Constraints | Other informations

Observatory BVET

General informations

Title	Observations and measurements made by observatory BVET
Abstract	
Identifier(s)	Empty list
Status	

Links

Internet links	Empty list
Snapshot(s)	Empty list

Observatory metadata consultation

DATA USERS

BVET

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search
- Advanced search

Identification | Contacts | Keywords | Geographical Informations | Temporal informations | Constraints | Other informations

Observatory BVET

General informations

Title	Observations and measurements made by observatory BVET
Abstract	
Identifier(s)	Empty list
Status	

Experimental site metadata consultation

DATA USERS

BVET > MENGONG

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search
- Advanced search

Identification | Contacts | Keywords | Geographical Informations | Temporal informations | Constraints | Other informations

Experimental site

Name	MENGONG

General informations

Title	Observations and measurements made by experimental site MENGONG
Abstract	Abstract EN
Identifier(s)	Empty list
Status	

Links

Internet links	Empty list
Snapshot(s)	Empty list

Observatory metadata consultation

DATA USERS | BVET

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search
- Advanced search

Identification | Contacts | Keywords | Geographical Informations | Temporal informations | Constraints | Other informations

Observatory BVET

General informations

Title	Observations and measurements made by observatory BVET
Abstract	
Identifier(s)	Empty list
Status	

Experimental site metadata consultation

DATA USERS | BVET | MENGONG

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search

Identification | Contacts | Keywords | Geographical Informations | Temporal informations | Constraints | Other informations

Experimental site

Name	MENGONG

General informations

Title	Observations and measurements made by experimental site MENGONG

Dataset metadata consultation

DATA USERS | BVET | MENGONG

- Browse catalog
  - Observatories
  - Climates
  - Geologies
- Search
- Advanced search

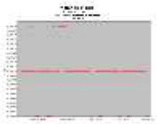
Identification | Contacts | Keywords | Geographical Informations | Temporal informations | Constraints | Other informations

General informations

Title	BVET - Hydrological data - Nsimi (spring), Cameroon
Abstract	Hydrological data collected at the Nsimi (spring) site (Mengong hydrosystem, Cameroon).
Identifier(s)	- 10.6096/BVET.CMR.HYDRO (http://dx.doi.org/)
Status	

Links

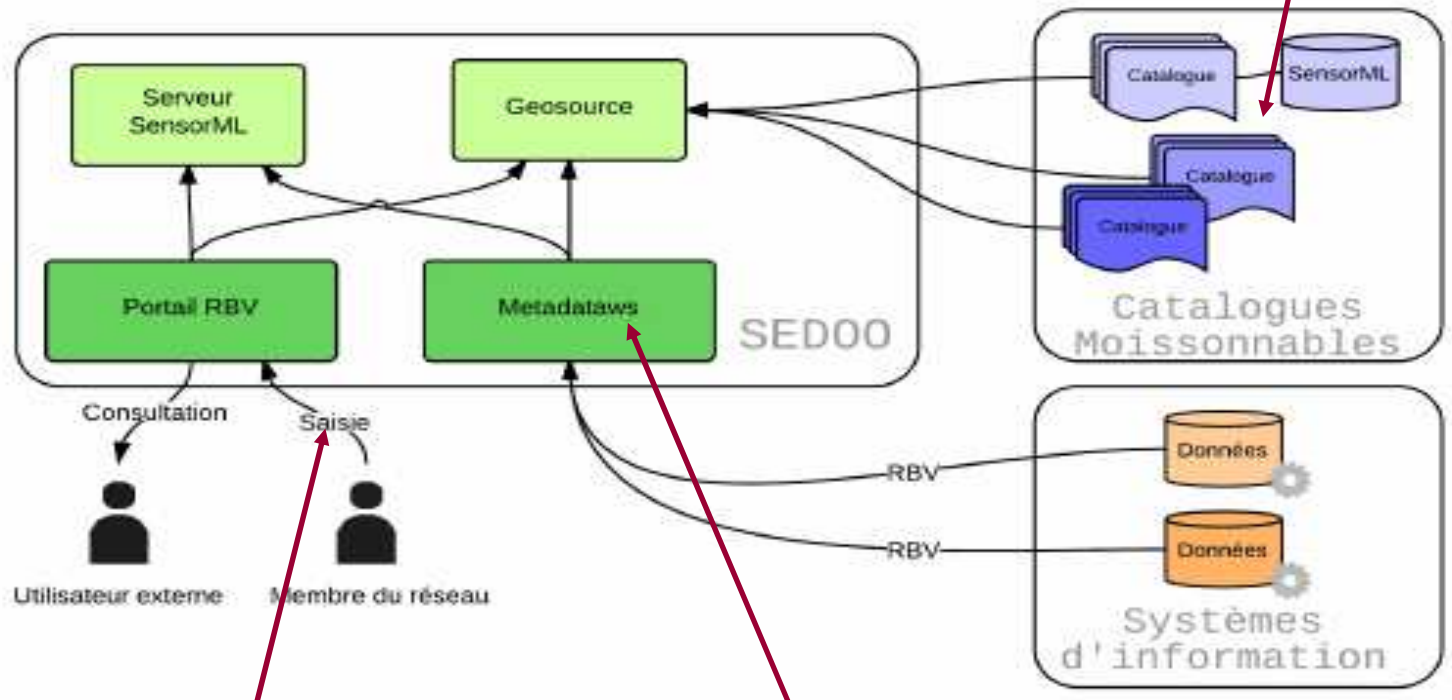
Internet links	- Water discharge INFORMATION
Snapshot(s)	- Water discharge



Automatic information transfer

RBV metadata files feeding ways

Automatic metadata harvesting:
catalogues synchronized daily



Online form filling

Web service generating ISO 19139
profile from a simplified RBV profile

Data homogenization

Work to be done by the data providers or the data managers?

- **Case 1: data providers agree on a common data format**. Disciplinary « **standard » format** is recommended (NASA-Ames, NetCDF…)

- **Case 2: insertion of all the datasets in a relational database (1D data) or processing/conversion of all the gridded data by the data managers**.

  - Needs lot of manpower. Errors insertion are possible.

  + Easy-to-use data for the user. Complex data requests possible (e.g. ‘’select all the observations between date 1 and date 2, in a cube of space, where parameter λ exceeds a certain value’’).

- **Case 3:** HyMeX and ChArMEx strategy. **Processing level defined for every dataset** following cost / advantages evaluation. In general datasets are inserted if they will be widely used, if they will probably be used by non-expert users or if a same instrument or property is provided in many different formats.

Data search

Keywords :	synop	
Search with:	All of the above keywords / Any of the above keywords	
Instrument type		
Parameter keyword	-- Topic -- / -- Term -- / -- Var_level1 -- / -- Var_level2 --	
Period	Date begin (yyyy-mm-dd)	Date end (yyyy-mm-dd)
Sort result:	By instruments / By platform types	

Zone

UnZoom

Zoom

Lat: 45 ° 0 ' 0 "
Lon: 4 ° 52 ' 29 "

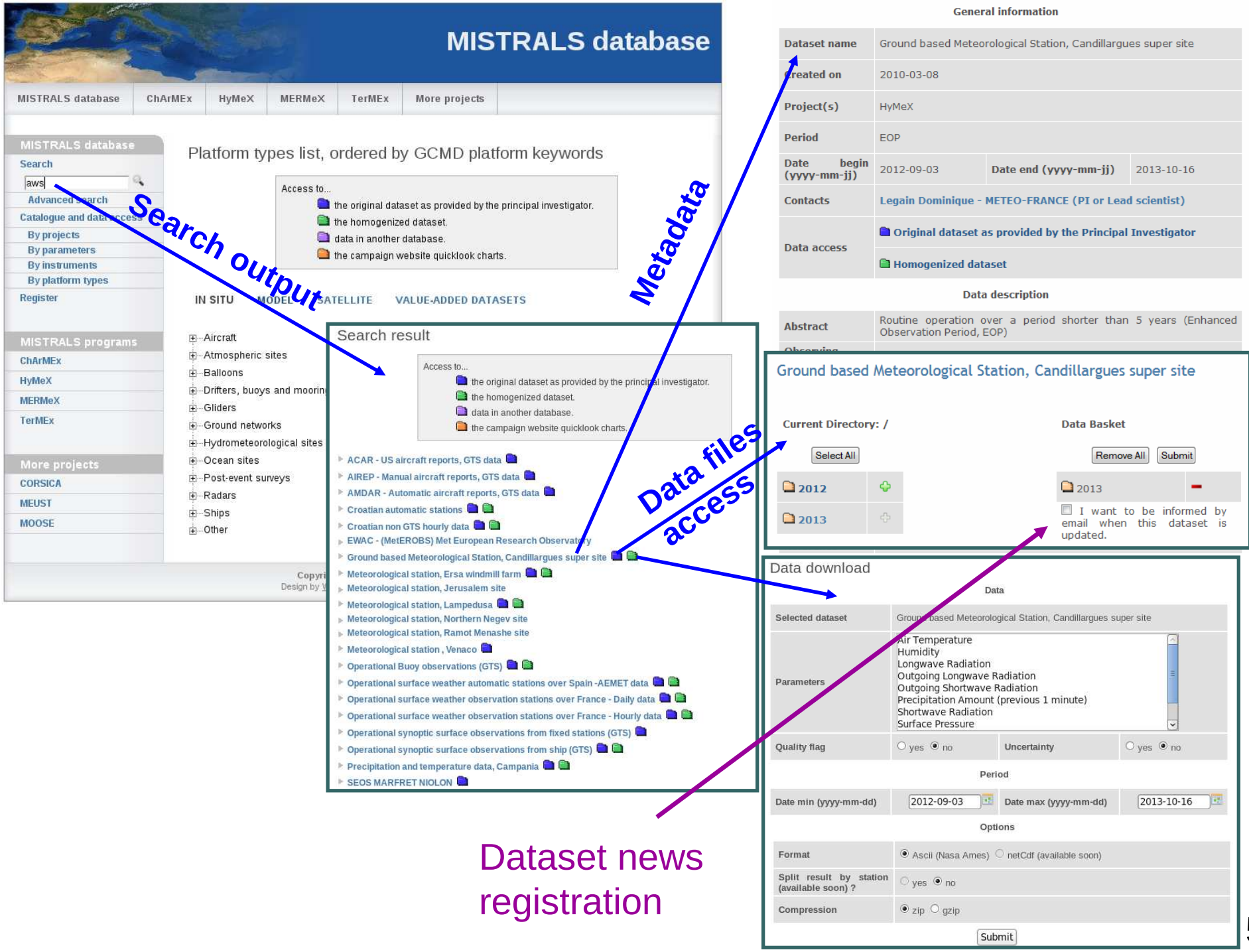
Lat: 38 ° 52 ' 29 "
Lon: 13 ° 52 ' 29 "

Show only datasets with available data ?	yes / no	
Show only datasets with homogenized data ?	yes / no	

search

Metadata search tool and data access

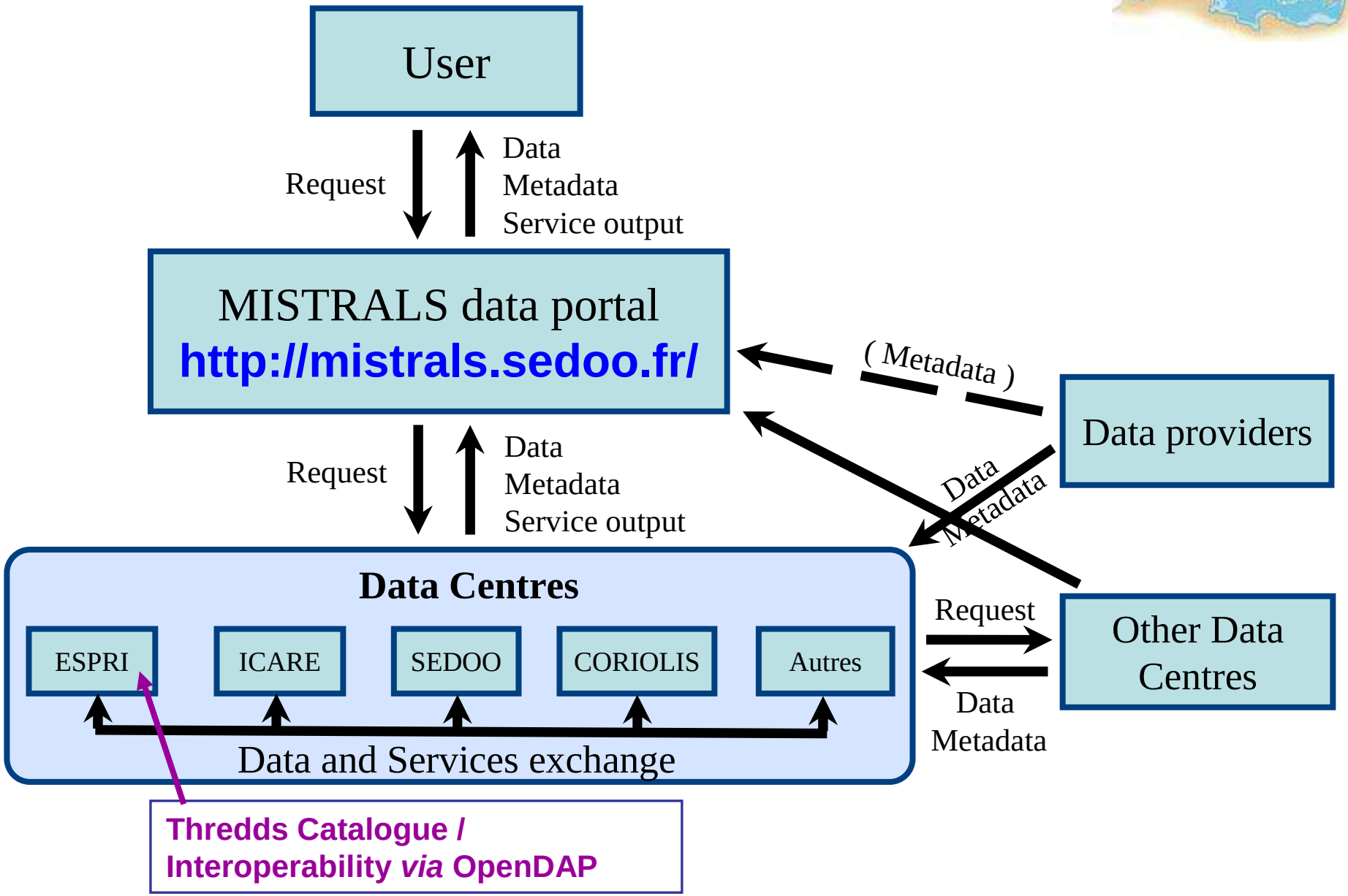
Cross-criteria or keyword search


MISTRALS database
MISTRALS database
ChArMEx
HyMeX
MERMeX
TerMEx
More projects
MISTRALS database
Search
aws
Advanced search
Catalogue and data access
By projects
By parameters
By instruments
By platform types
Register
MISTRALS programs
ChArMEx
HyMeX
MERMeX
TerMEx
More projects
CORSICA
MEUST
MOOSE
Platform types list, ordered by GCMD platform keywords
Access to...
the original dataset as provided by the principal investigator.
the homogenized dataset.
data in another database.
the campaign website quicklook charts.
IN SITU
MODEL
SATELLITE
VALUE-ADDED DATASETS
Aircraft
Atmospheric sites
Balloons
Drifters, buoys and mooring
Gliders
Ground networks
Hydrometeorological sites
Ocean sites
Post-event surveys
Radars
Ships
Other
Search output
Search result
Access to...
the original dataset as provided by the principal investigator.
the homogenized dataset.
data in another database.
the campaign website quicklook charts.
ACAR - US aircraft reports, GTS data
AIREP - Manual aircraft reports, GTS data
AMDAR - Automatic aircraft reports, GTS data
Croatian automatic stations
Croatian non GTS hourly data
EWAC - (MetEROBS) Met European Research Observatory
Ground based Meteorological Station, Candillargues super site
Meteorological station, Ersa windmill farm
Meteorological station, Jerusalem site
Meteorological station, Lampedusa
Meteorological station, Northern Negev site
Meteorological station, Ramot Menashe site
Meteorological station , Venaco
Operational Buoy observations (GTS)
Operational surface weather automatic stations over Spain -AEMET data
Operational surface weather observation stations over France - Daily data
Operational surface weather observation stations over France - Hourly data
Operational synoptic surface observations from fixed stations (GTS)
Operational synoptic surface observations from ship (GTS)
Precipitation and temperature data, Campania
SEOS MARFRET NIOLON
Metadata
General information
Dataset name
Ground based Meteorological Station, Candillargues super site
Created on
2010-03-08
Project(s)
HyMeX
Period
EOP
Date begin (yyyy-mm-jj)
2012-09-03
Date end (yyyy-mm-jj)
2013-10-16
Contacts
Legain Dominique - METEO-FRANCE (PI or Lead scientist)
Data access
Original dataset as provided by the Principal Investigator
Homogenized dataset
Data description
Abstract
Routine operation over a period shorter than 5 years (Enhanced Observation Period, EOP)
Data files access
Ground based Meteorological Station, Candillargues super site
Current Directory: /
Select All
2012
2013
Data Basket
Remove All
Submit
2013
I want to be informed by email when this dataset is updated.
Data download
Data
Selected dataset
Ground based Meteorological Station, Candillargues super site
Parameters
Air Temperature
Humidity
Longwave Radiation
Outgoing Longwave Radiation
Outgoing Shortwave Radiation
Precipitation Amount (previous 1 minute)
Shortwave Radiation
Surface Pressure
Quality flag
yes
no
Uncertainty
yes
no
Period
Date min (yyyy-mm-dd)
2012-09-03
Date max (yyyy-mm-dd)
2013-10-16
Options
Format
Ascii (Nasa Ames)
netCdf (available soon)
Split result by station (available soon) ?
yes
no
Compression
zip
gzip
Submit
Dataset news registration

Interoperability *vs* duplication

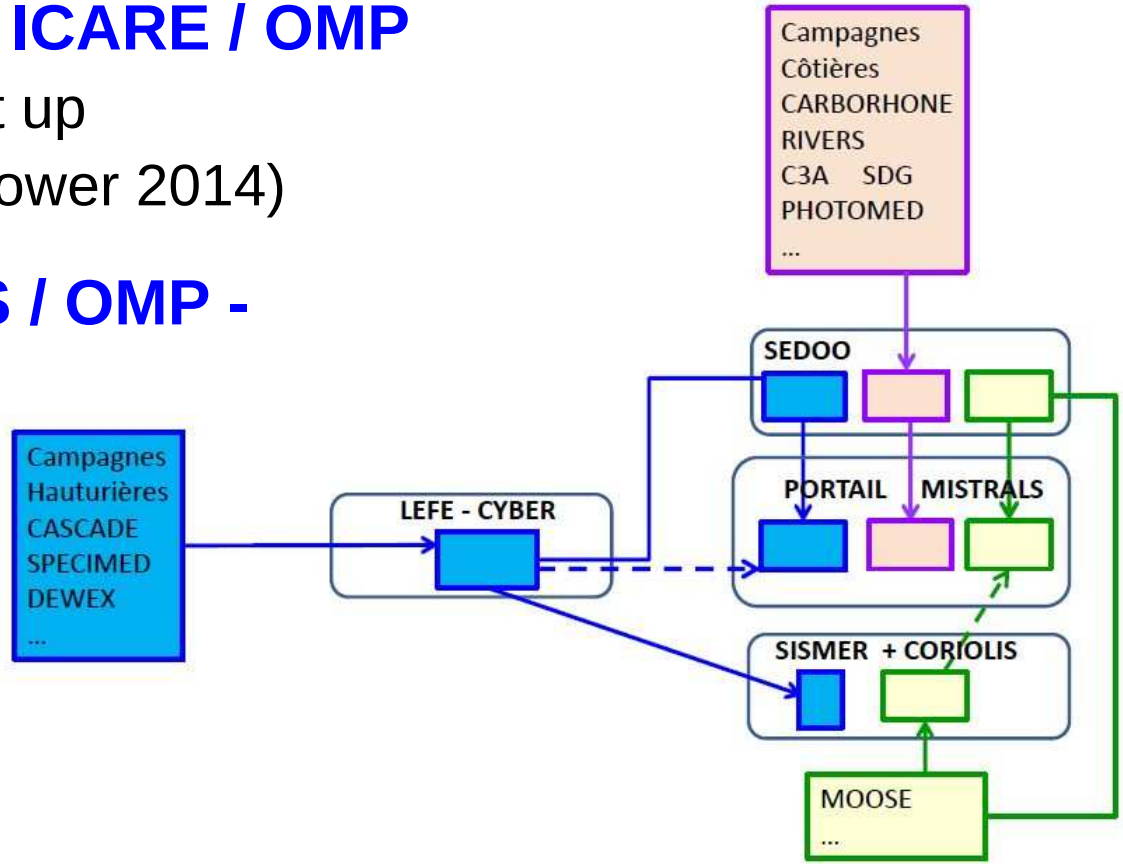
- **Complete interoperability**
  - search, select and access to every dataset in different data centres from the single data portal
  - user request results are synchronized
  - Problem: requires development / manpower in the different data centres
  - May be slow or not work depending of the data centres operational status
- **Duplication**
  - relevant if low data volume
  - **identify the primary data center** communicating with the data provider, then sending data to the other data centers.
- **Other options**
  - Metadata containing data access procedure description
  - Metadata containing the exact dataset url link
  - Idem + same user login/pass in the portal and the external data centre
  - Dataset request sent to the external data centres. The user will receive a request output by every data centre.

Architecture



Communication protocols

- **Interoperability IPSL / OMP – IPSL / ENEA**
  - at present: user directory synchronization + ftp protocol
- **Interoperability IPSL / ICARE / OMP**
  - protocol OpenDAP set up
  - test to be done (manpower 2014)
- **Duplication CORIOLIS / OMP - LEFE-CYBER / OMP**



Data policy agreement purposes

- Data policy is a key issue for an **efficient data exchange** inside the project and beyond it.

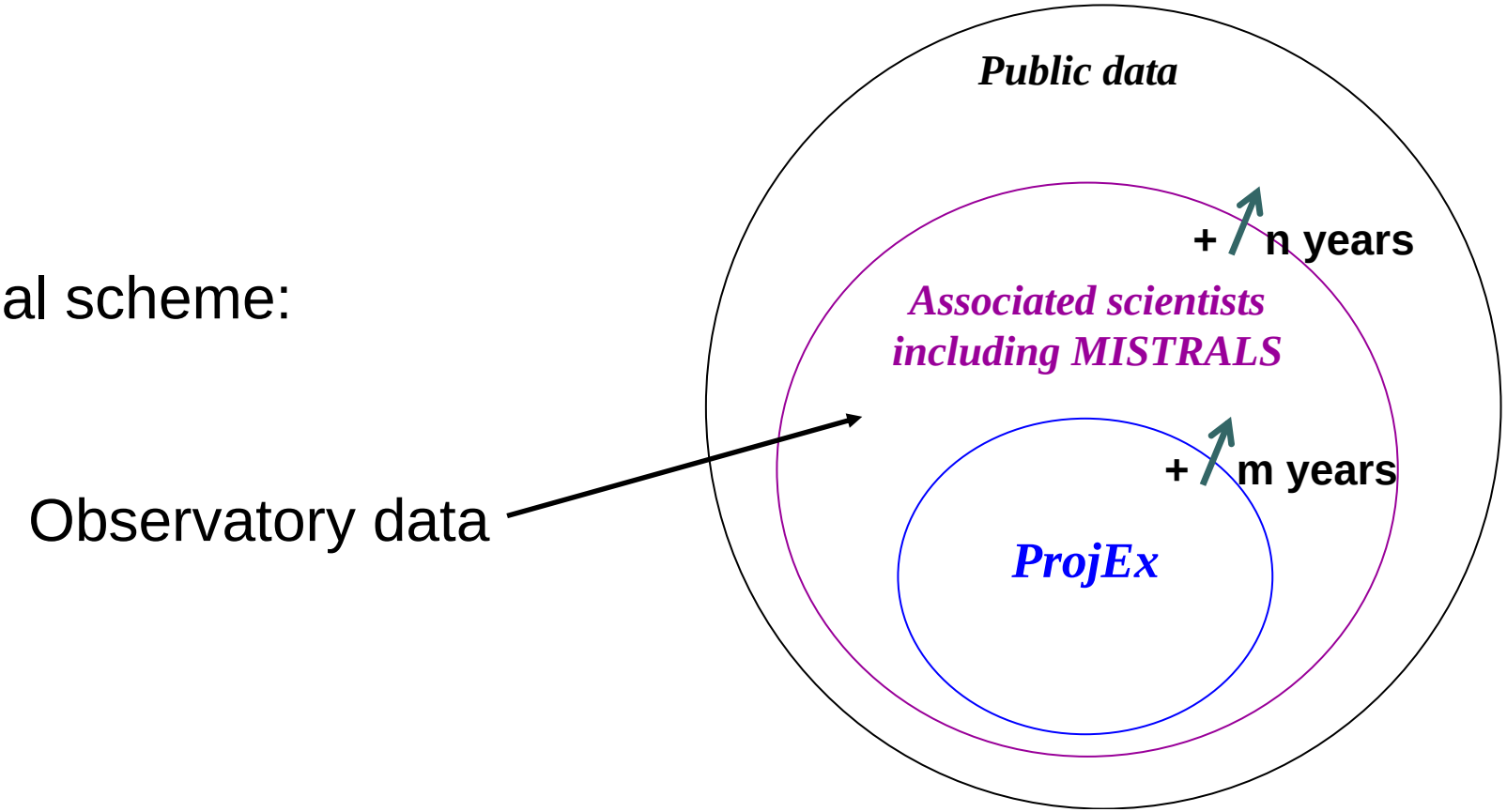
- Defining **mutual duties and rights of data producers and data users**.
  - Data users need: - « good quality » data « early enough ».
    - data producer's expertise.
  - Data producers need: - to participate in or at least be aware of data users activities.
    - to co-sign or at least be acknowledged in publications involving the dataset they produced.

- Data acquisition **funding project or institution** need to be acknowledged in publications / contributions.

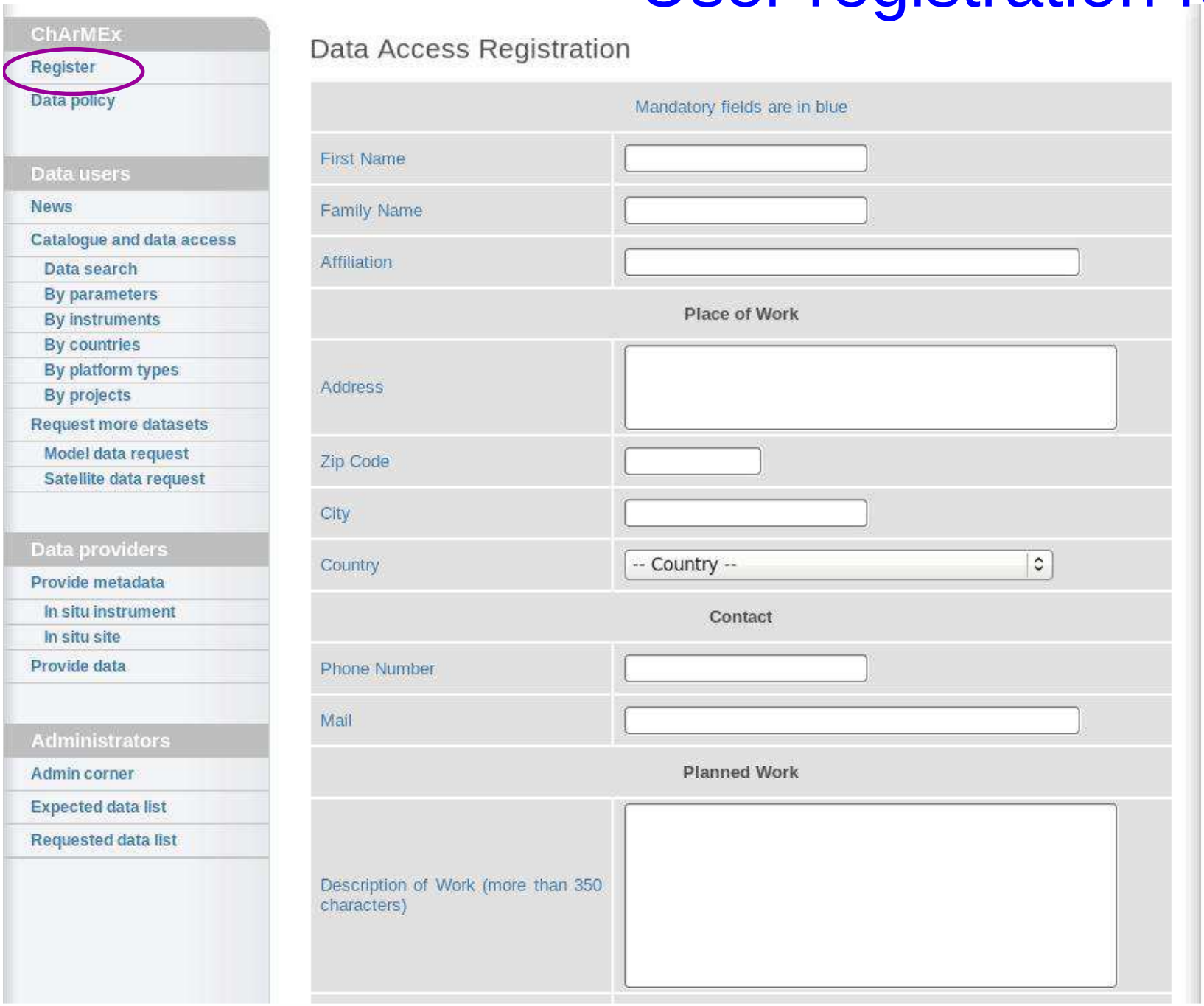
MISTRALS project data policy

- Global agreement on the « good behaviour » rules
- But: different user / data access circles definition following the project, with different reserved period.

Typical scheme:



User registration form



Data policy

With this form, I fully accept the ChArMEx data and publication policy (**click here to access pdf file**). In particular:

- ☐ I apply for access to the data in order to perform the work described above.
- ☐ I commit not to use data for commercial applications.
- ☐ I commit to never redistribute data to any other person nor institution.
- ☐ I commit to contact the provider of the required data in order to offer collaboration.

automatic emails

In the case of any publication or communication using datasets or products obtained in the database:

- ☐ I commit to agree on either co-authorship or acknowledgement with the principal investigators of the datasets.
- ☐ I commit to acknowledge the agencies which have funded the data collection and the database centres, using the mentions indicated in the dataset metadata and files.
- ☐ I commit to send the publication to the ChArMEx Project Office using the address **publicharmex@sedoo.fr** as soon as it is submitted to a journal, conference or other organisation.

If I process a dataset or product based on data received under the ChArMEx data policy:

- ☐ I commit to inform the database managers.
- ☐ I accept that the processed dataset or product may be disseminated under the ChArMEx data policy, if so decided by the ChArMEx SC.

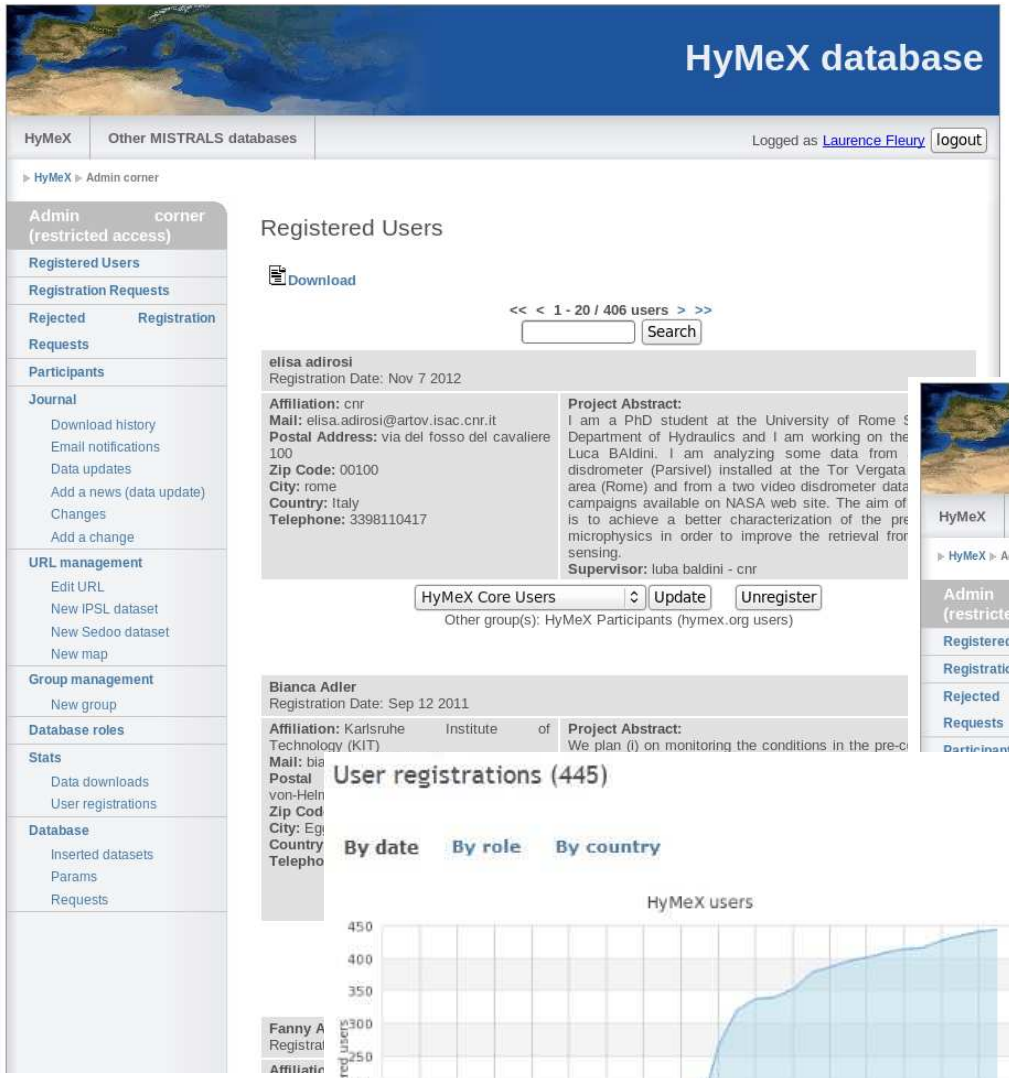
Apply

Validation by the project leader

Specific acknowledgement sentence is part of the dataset metadata and sent to the data user

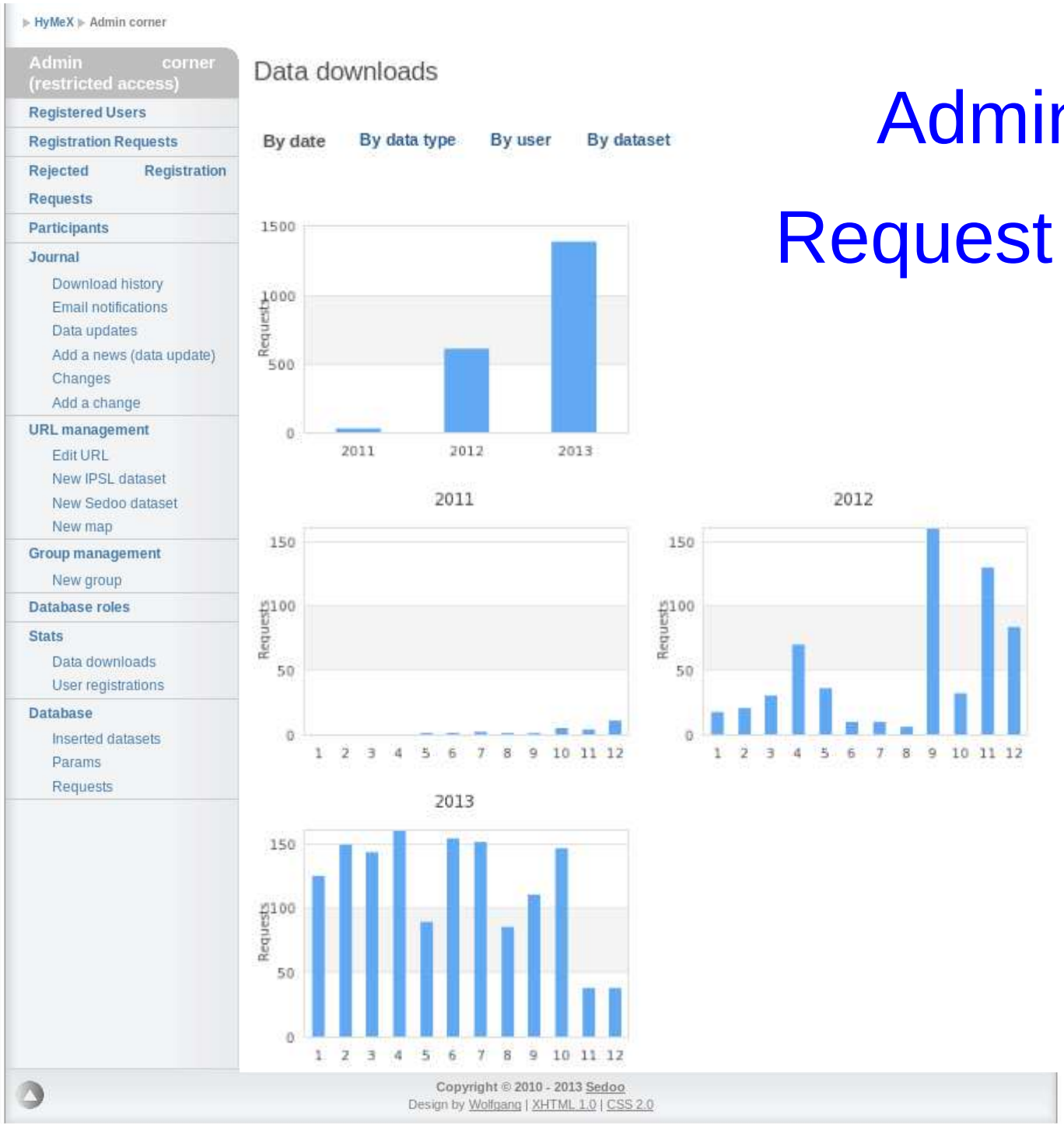
Admin corner

User management



Admin corner

Request monitoring



Thank you!

Contact: laurence.fleury@obs-mip.fr