A generic database for tagging data and much more

Julien Barde, Norbert Billet

[1]Institut de Recherche pour le dévelopement, UMR EME 212, 65 avenue Jean Monnet, 34200 Sète, France

30-10-2012

Summary

Summary

Improving fisheries data management to focus on EAF

EAF requires data from various disciplines:

- different species (birds, mammals. . .) and observed characteristics:
  - fishery data (one source of ecological observations among others):
    - before EAF: mainly target species,
    - after: many species with different roles (by-cacth, preys. . .),
  - trophic data, contaminants, fatty acids. . . ,
  - . . . tagging data, telemetry (spaghetti, pop-up. . .) useful to calibrate models (growth. . .),
- . . . to be combined with environmental parameters (gridded data),
  - from sensors (remote sensing or in situ: satellite / aerial, CTD. . .),
  - from models.

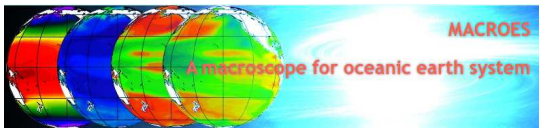
Motivation for a new database

- Current situation is a nightmare: too many databases (one per data type), heterogeneous and distributed (RDBMS, codes. . .),
- None of them enables to administrate all these kinds of data.

Set up a long term Information System (IS)

IRD projects related to management of tagging / EAF data:

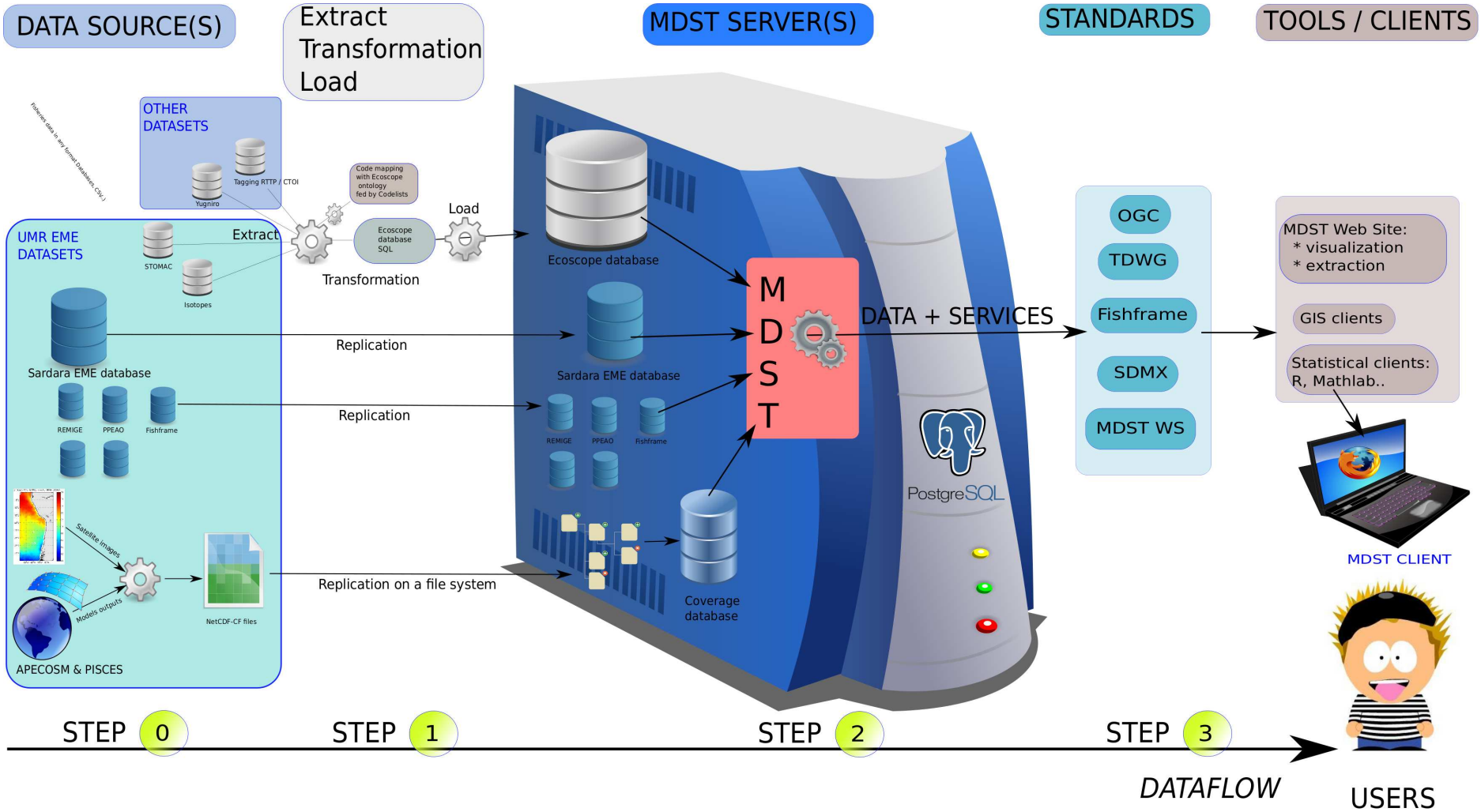
- REMIGE (ANR): 4 years (top predators),
- MACROES (ANR): 4 years (continuum of Remige, climate change),
- iMarine (FP7): 30 months (interoperability),
- . . . and future project: EMOTION (ANR). . .



A single IS whatever the project: Ecoscope / MDST

- Goal: create a new database and IS to administrate as much data as possible within a single model / architecture.
- Reuse and improve it according to projects and fundings,
- Make a first version available through opensource licence

Overview: main steps



Summary

. . . means nothing but replication(s) of observations

Tagging = replication(s) of measures on same individuals collected in different contexts:

- (1): deployment of the tag (scientific context): limitated types of observations,
- (0..n): observations inbetween (depending on tags: tracking. . .),
- (1): final recovery (dead tags or individuals): many possible observations according to the context.

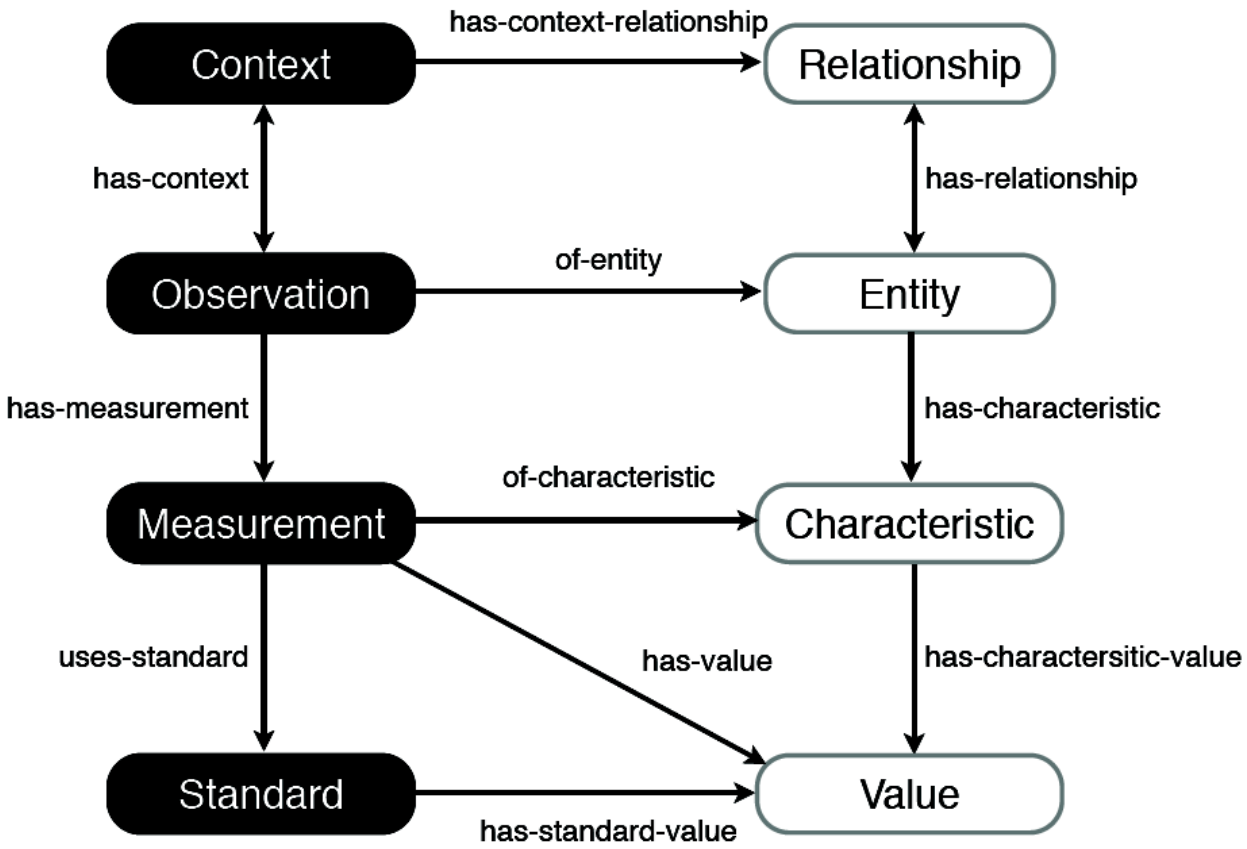
Tagging data = ?

- multiple kinds of (replicated) observations,
- multiple contexts for data collection: heterogeneity of quality.

Fisheries databases models don't fit the needs to manage tagging data: observations just made once.

Tagging & Ecological observations

Observation = "measurements directly linked to real-world phenomena".



[Bowers, 2008]

Tagging & Ecological observations

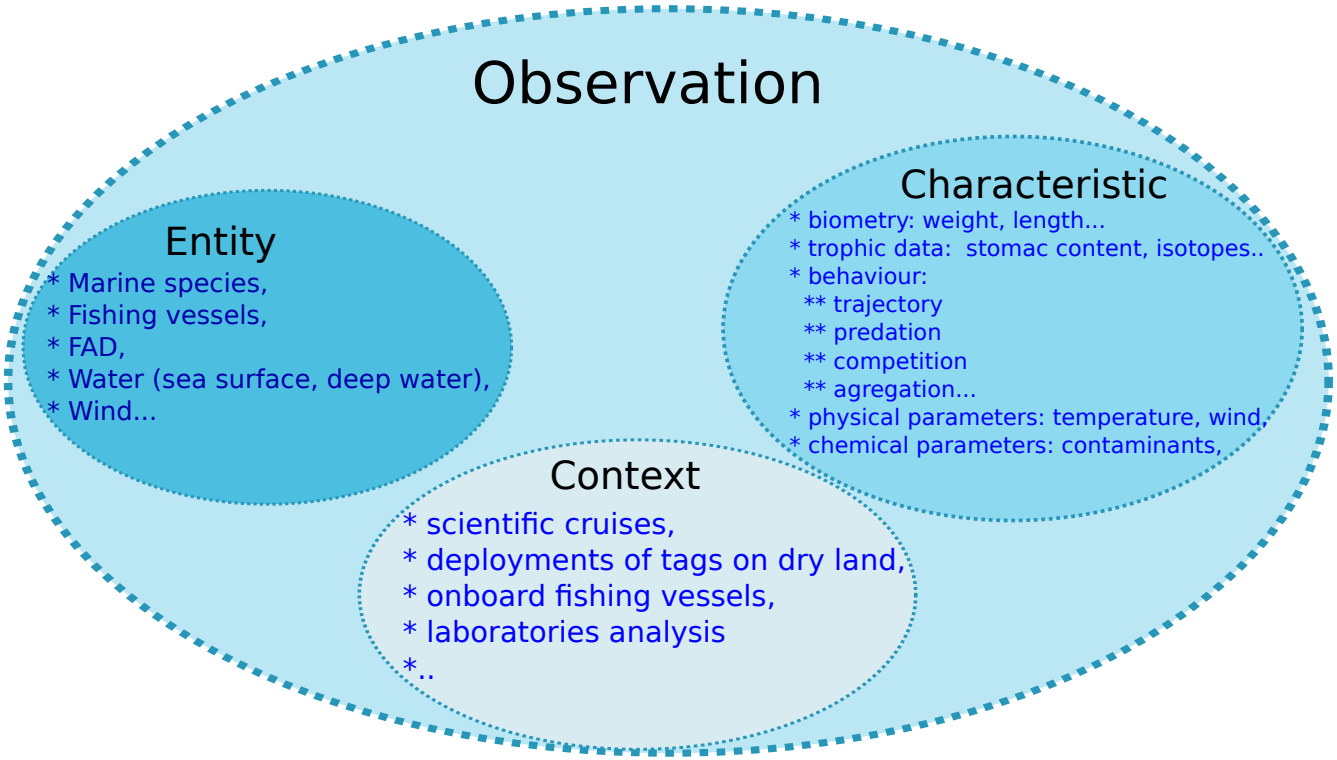
Observation = {characteristic of an entity, measure & value, context}

- entity? species, vessels, FAD, school, tissue samples. . . ,
- characteristic? weight, length, location. . . ,
- context? people/projects, location, date / period, sampling protocols,
- examples:
  - 1003={*activity of Kermantxo, tagging operation, 03/05/2006 11am*},
  - {*length of albacares 1234, 77 cm, 1003*},
  - 2434={*activity of Torre Gullia, fishing operation, 15/01/2007 5pm*},
  - {*weight of albacares 1234, 12.5 kg, 2434*},
  - {*length of albacares 1234, 82.8 cm, 2434*}.

Fisheries databases models = semantic traps:

- fishing effort *vs* observation effort, fishing gear *vs* sampling protocol,
- by-cacth / discards *vs* species / individuals,
- PS fishing operation *vs* school occurence. . .

EAF data & Ecological observations



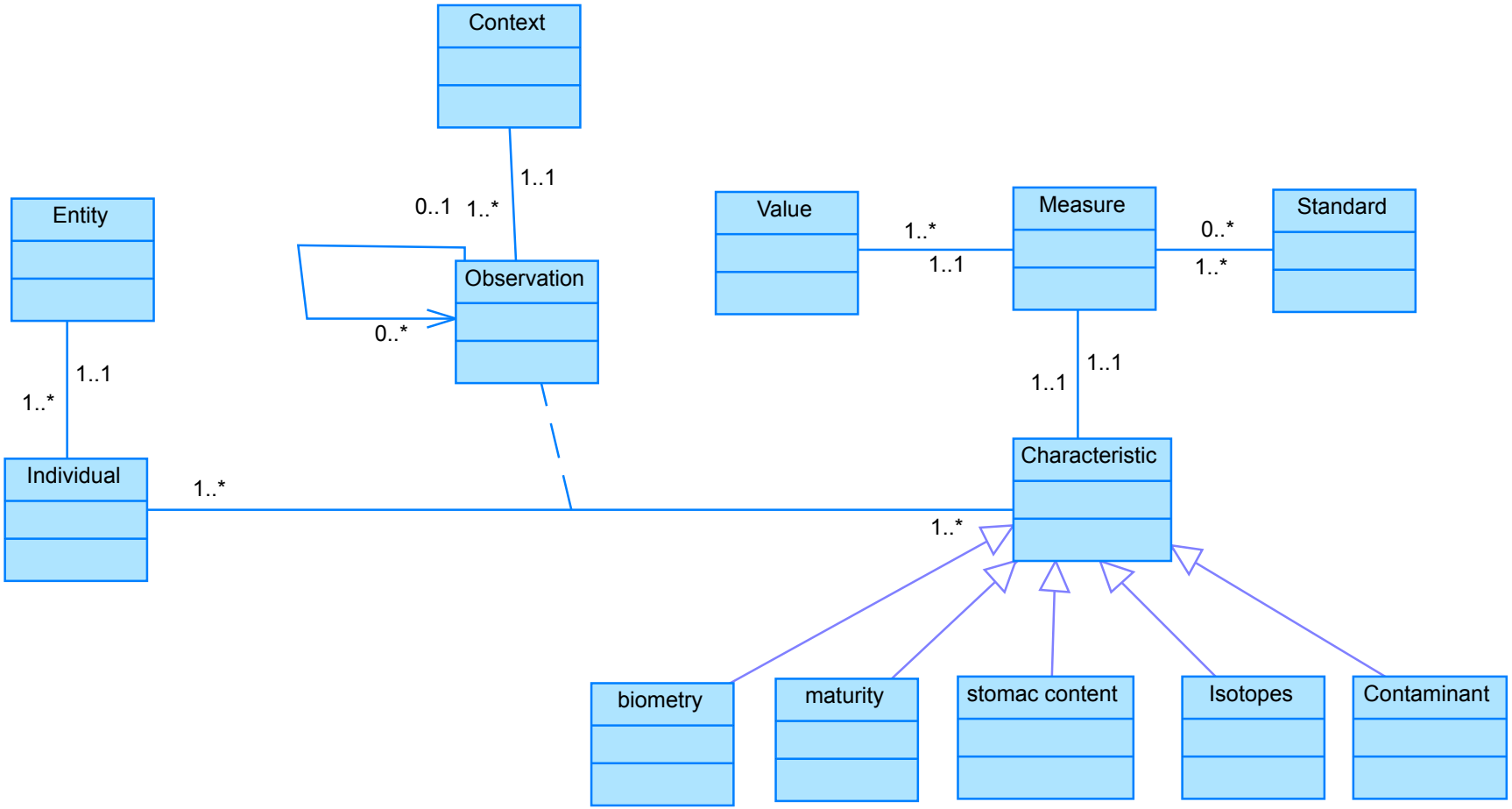
A generic database model for tagging data. . .

- . . . has to be a generic enough for various kinds of observations,
- . . . can be used to store EAF data (including fisheries data).

Summary

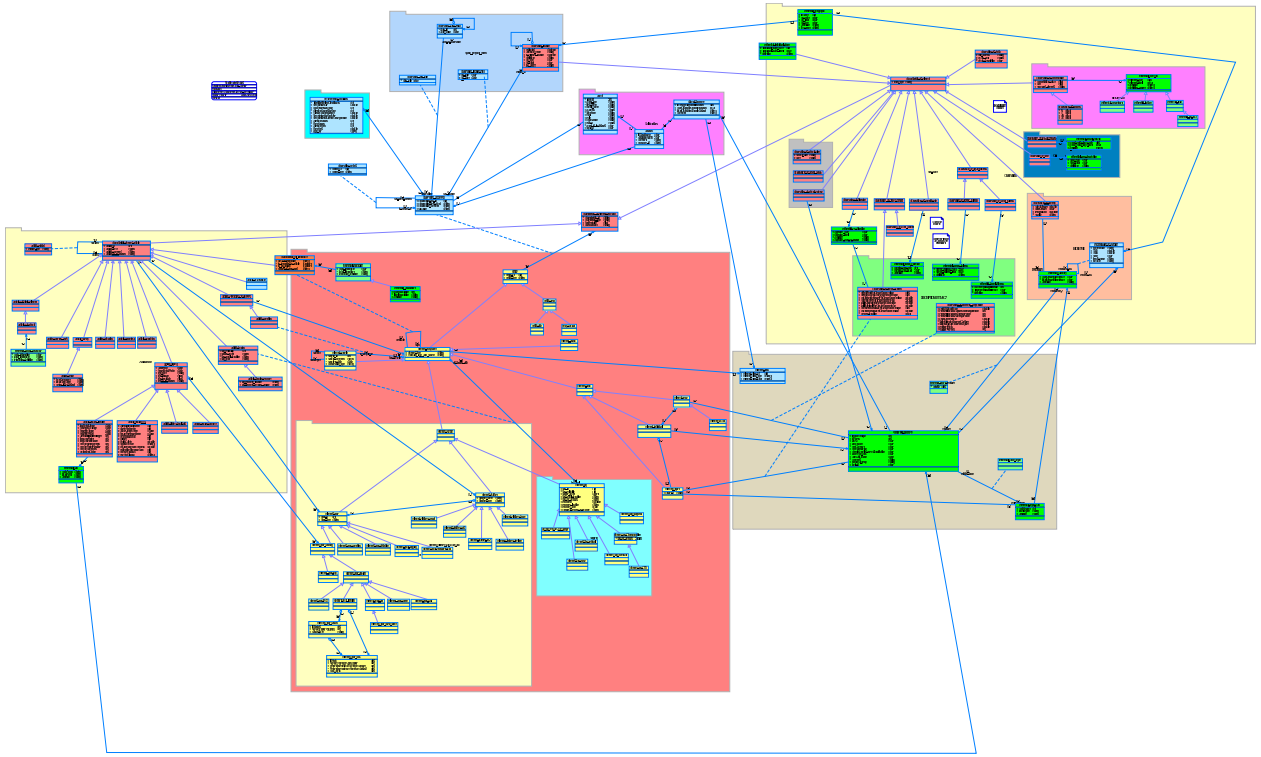
A new database model to store data (step 0)

It starts like this. . .



A new database model to store data (step 0)

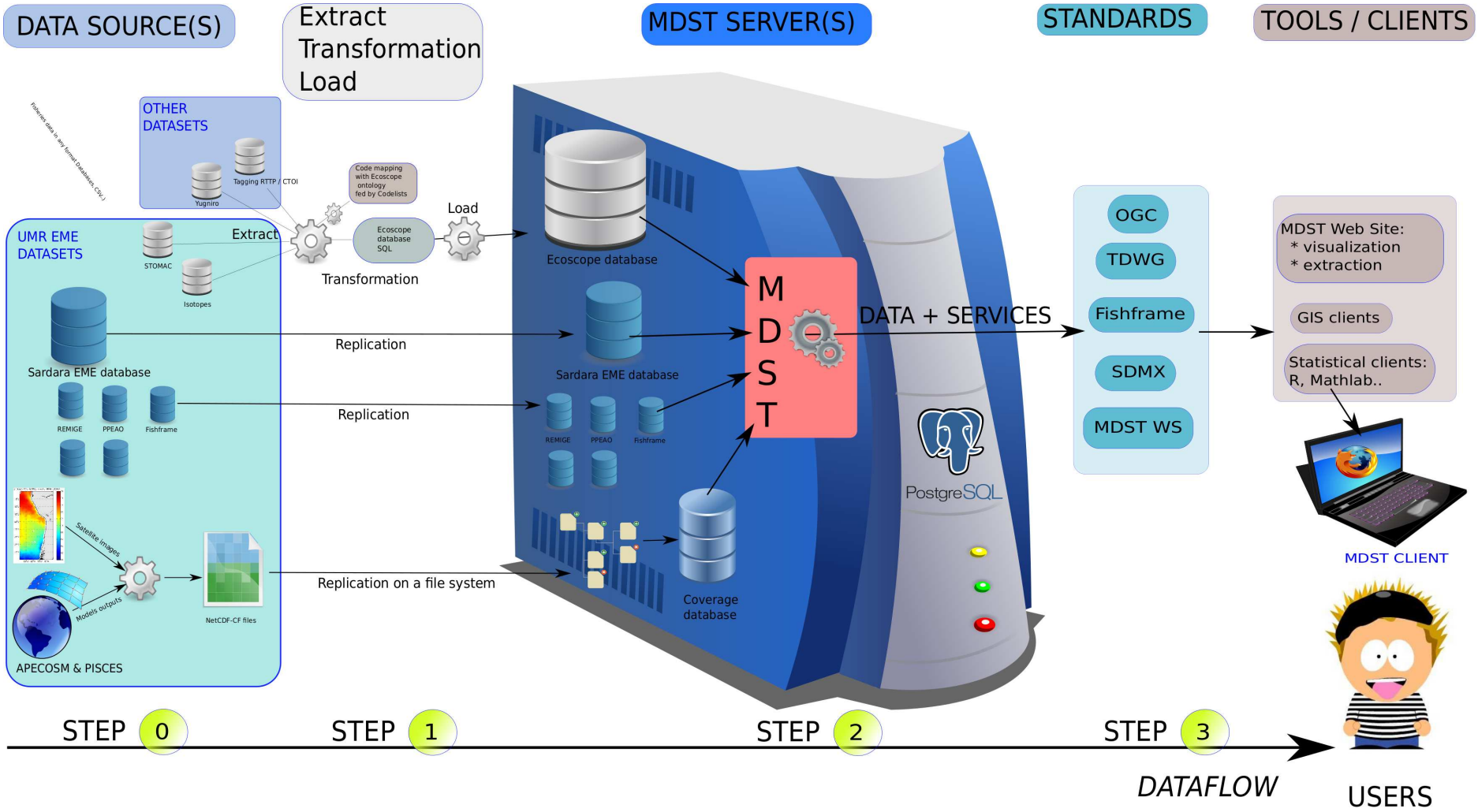
. . . and it ends up like this.



Implementation

- Postgres & Postgis 2.0,
- A set of common services for all data managed in this database. . .

Loading tagging data in the new database (step 1)



Loading tagging data in the new database (step 1)

Extract, Transform, Load (ETL), this step requires to::

- change codes (Worms. . .),
- change uom,
- change semantic of tables and attributes (columns),
- spatialize data.

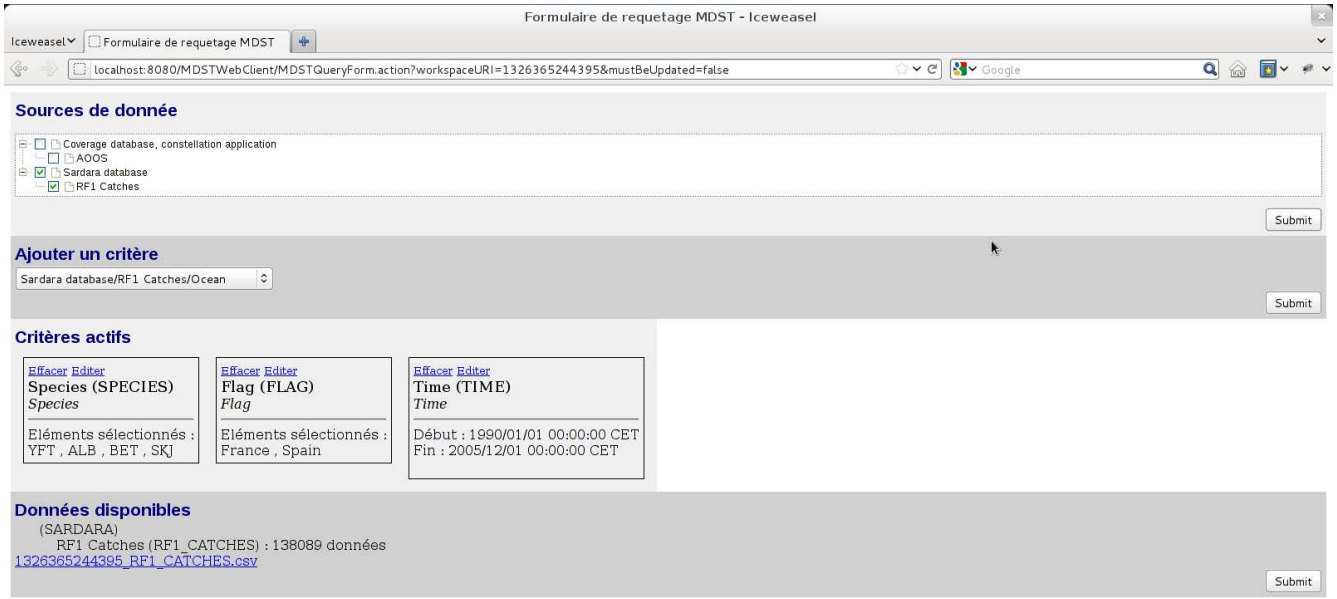
However the system can combine different approachs:

- transformation of data (a lot of work),
- no transformation (data is served as it is):
  - data stay in the data source,
  - conversion from access to postgres,
  - data types (spatialization. . .) and codes (worms. . .).

Dedicated data access on top of the database (step 2)

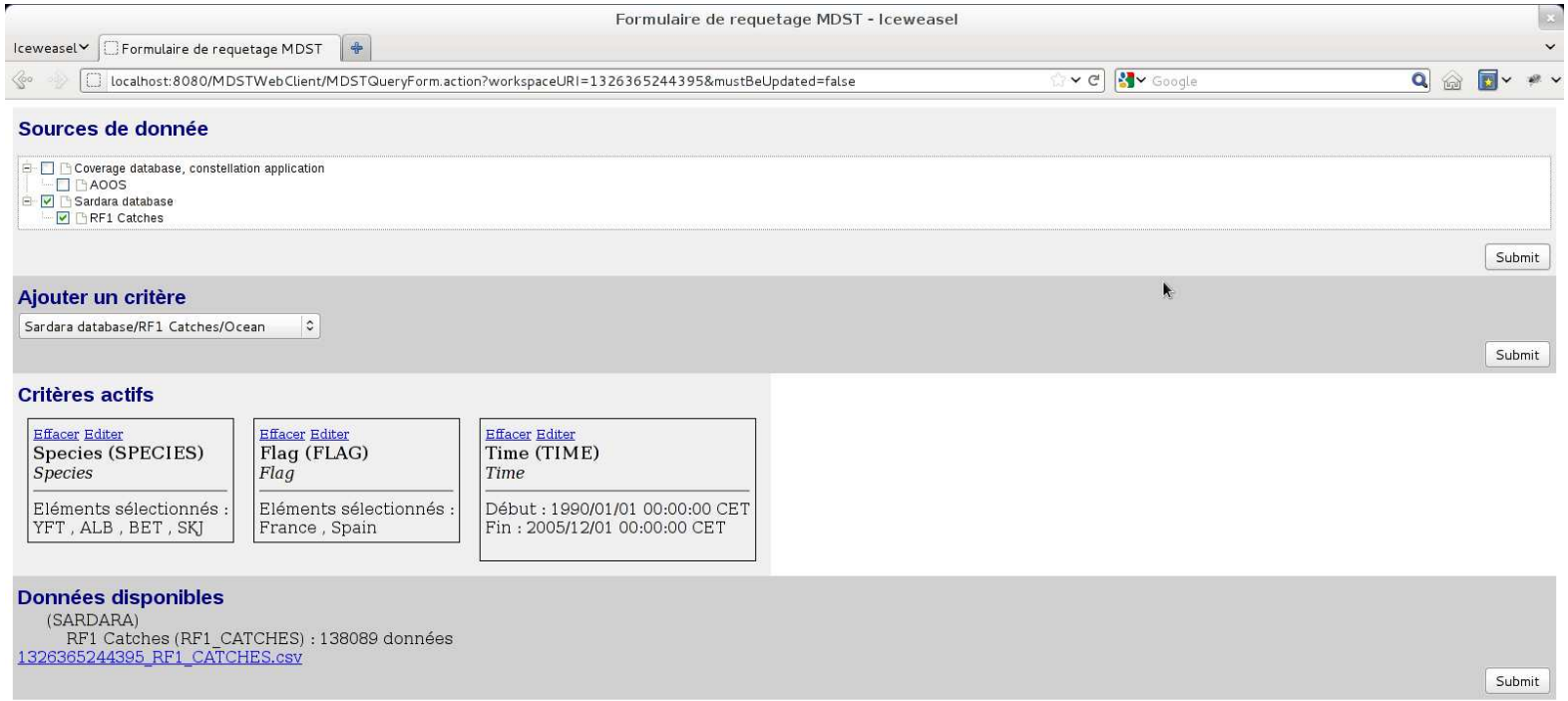
MACROES MDST server is a Web portal + other data access / services:

- the Web portal enables data access with a single html form whatever the number of data sources,
- . . . SQL queries,
- . . . Web Services,
- interoperability enables the data to be reused by other data sources.



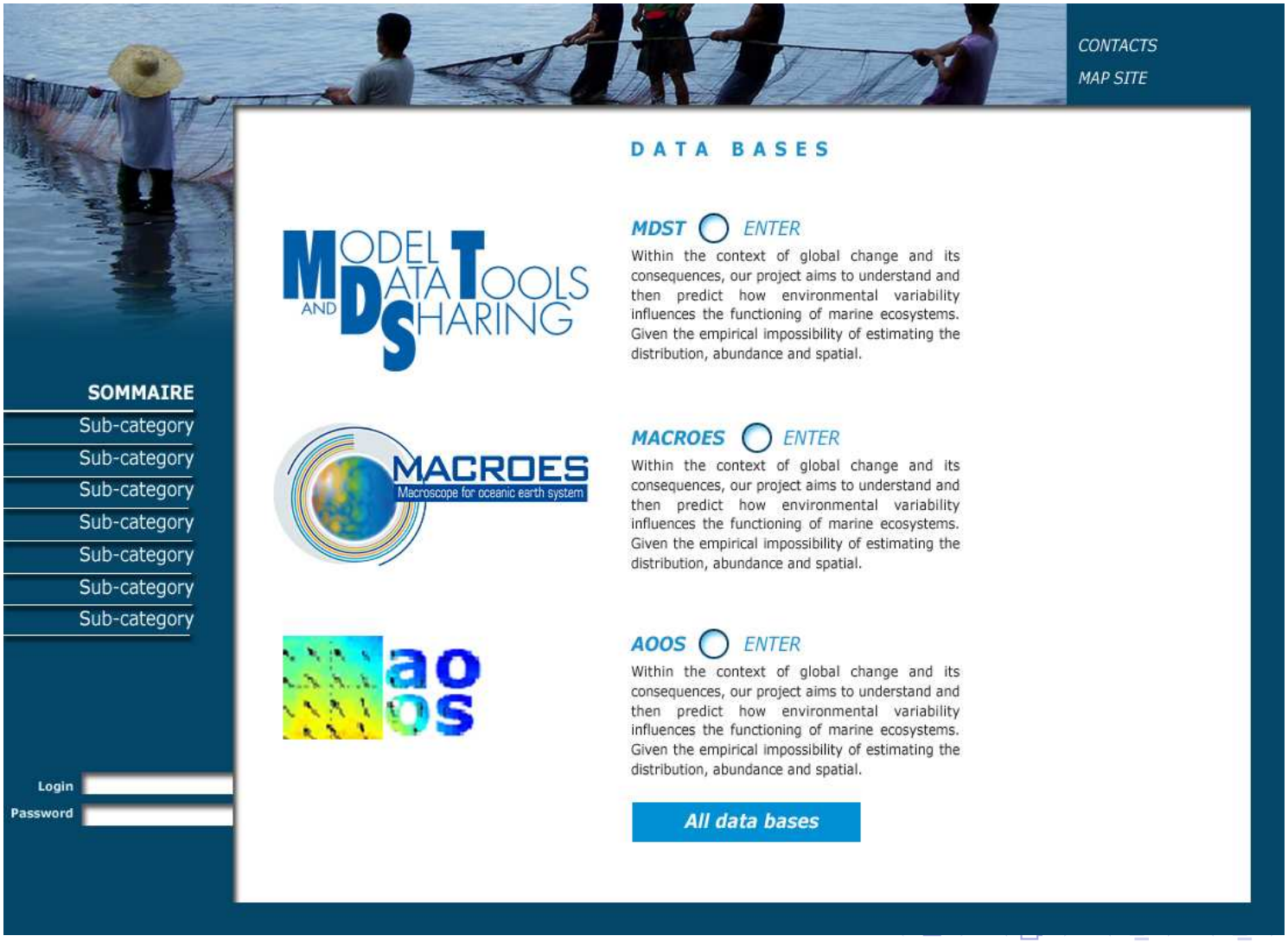
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



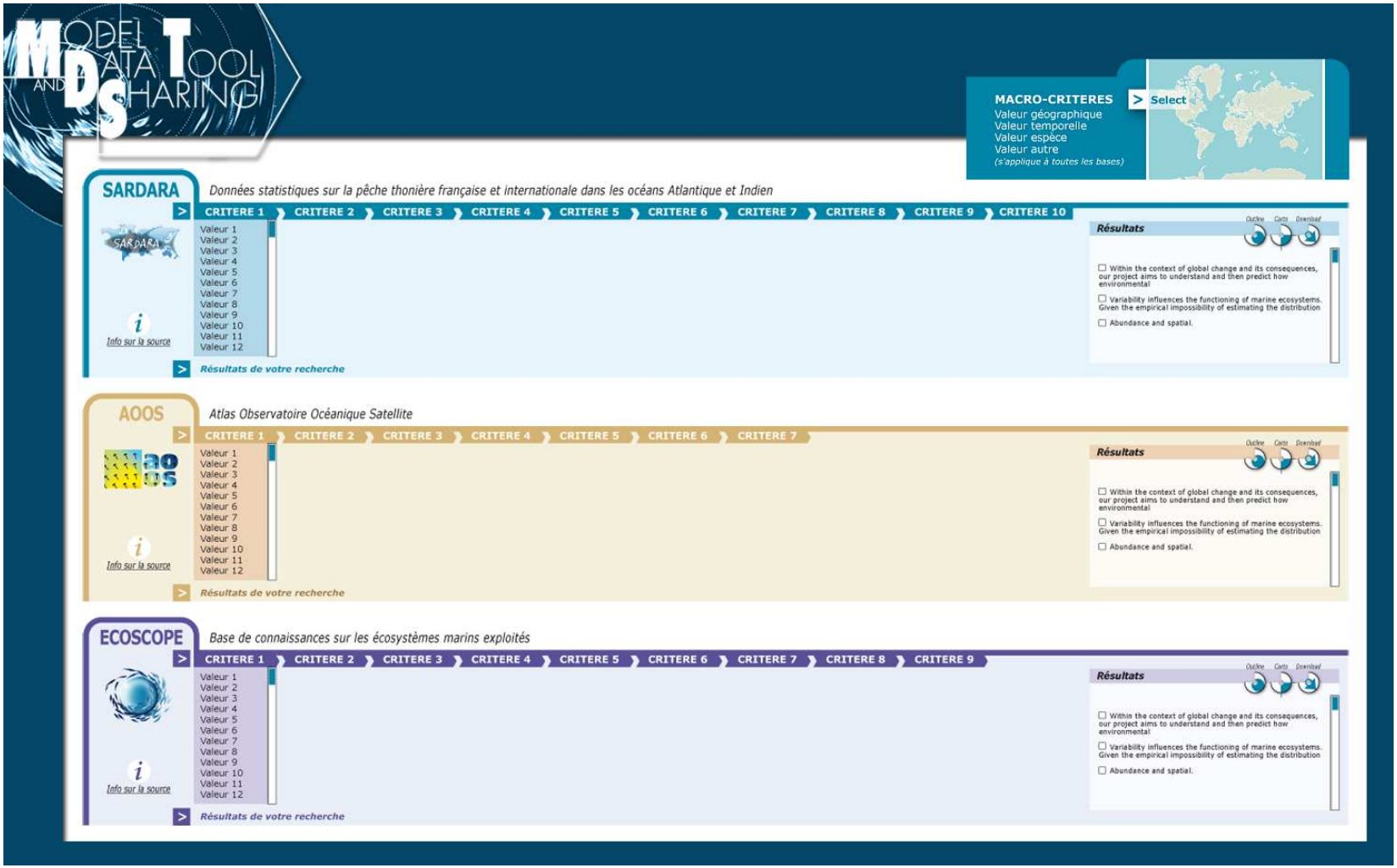
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



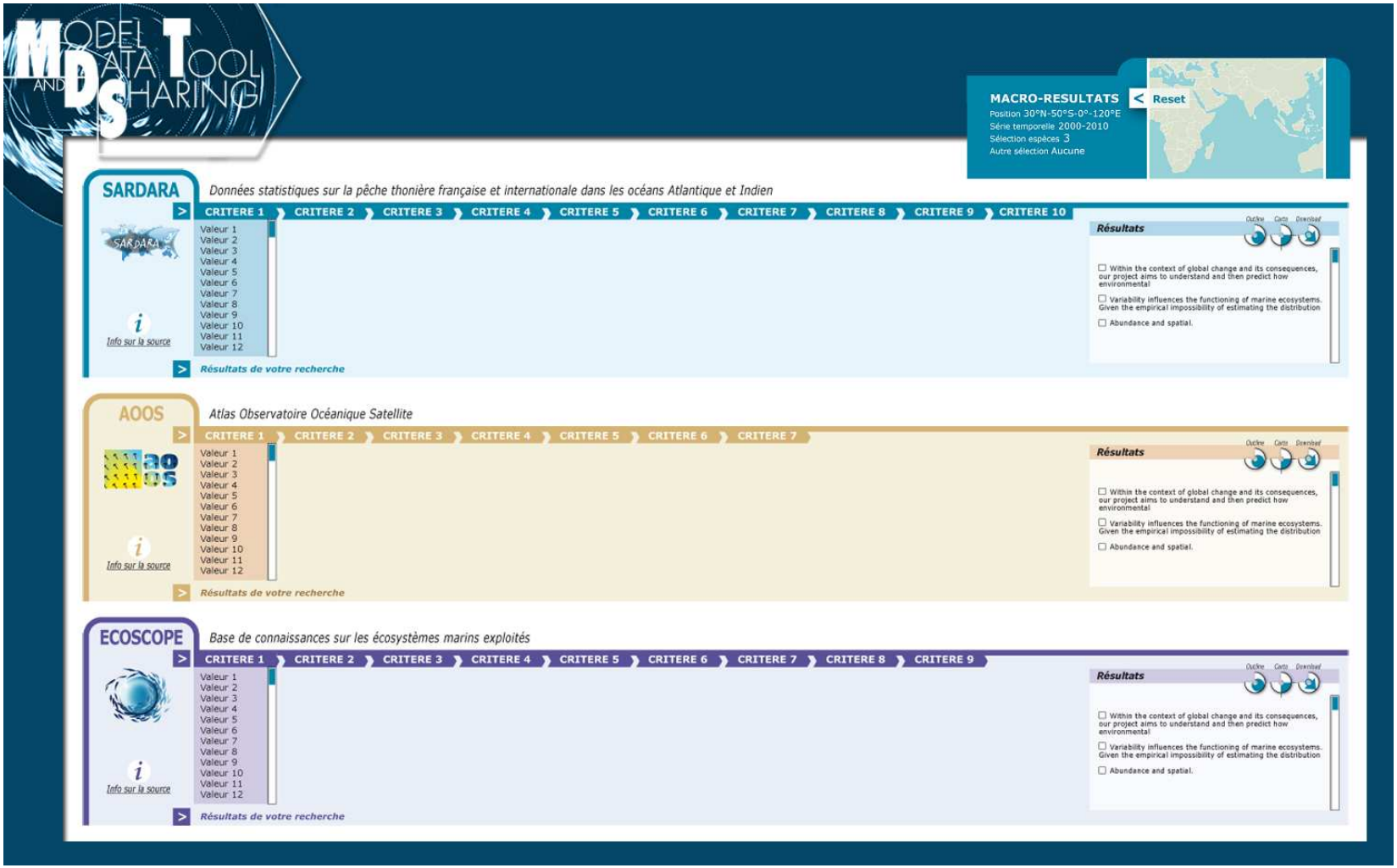
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



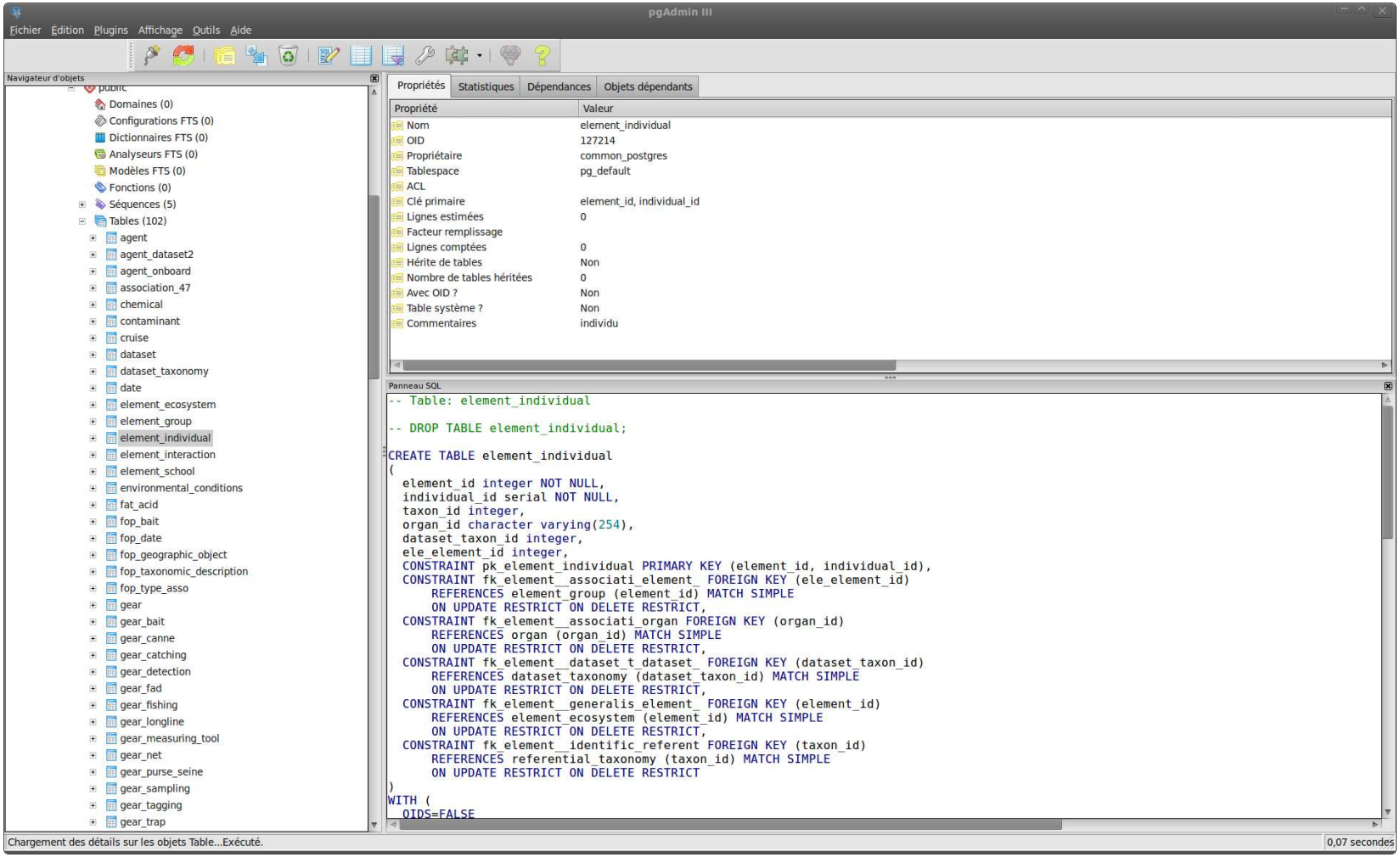
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



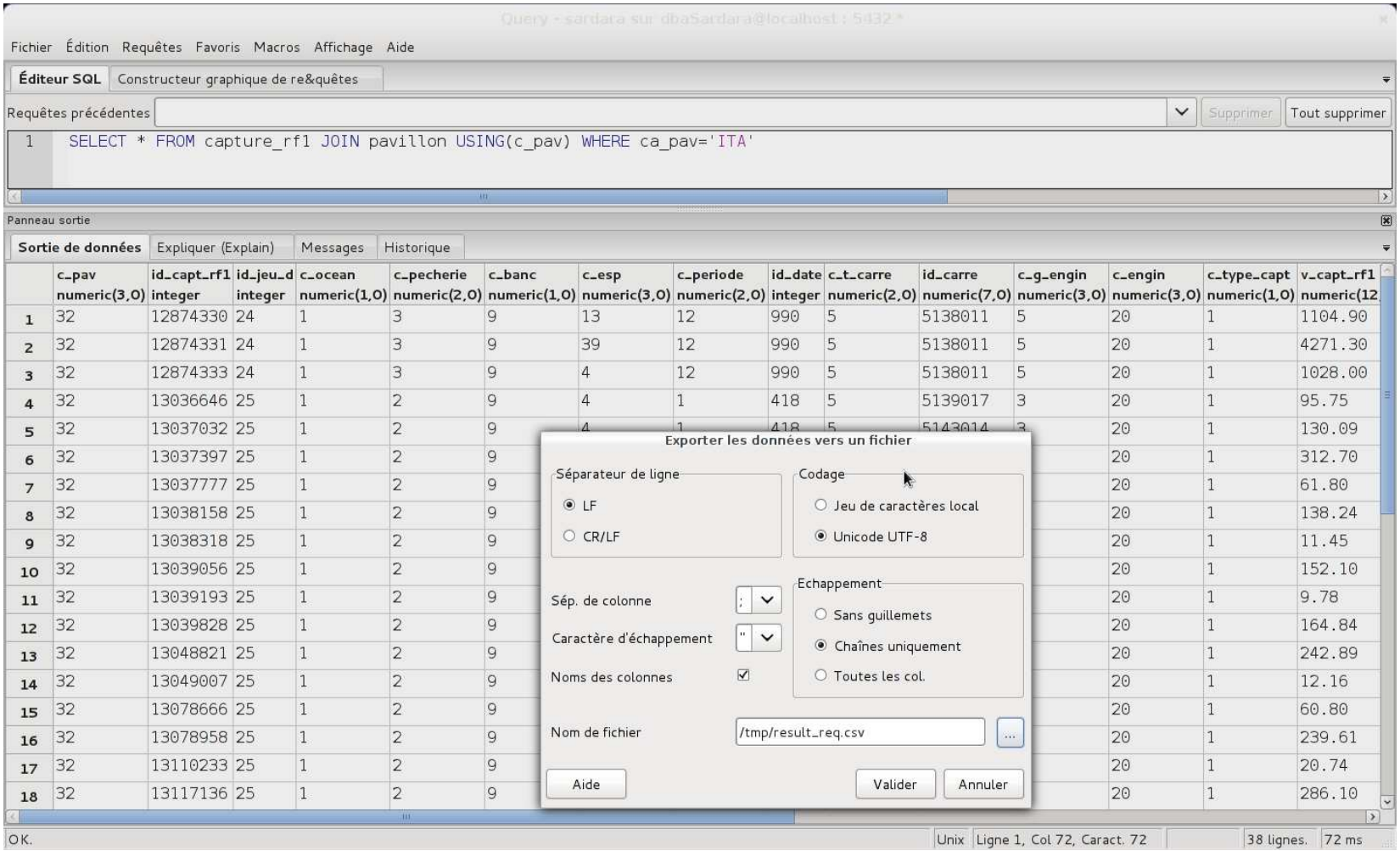
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



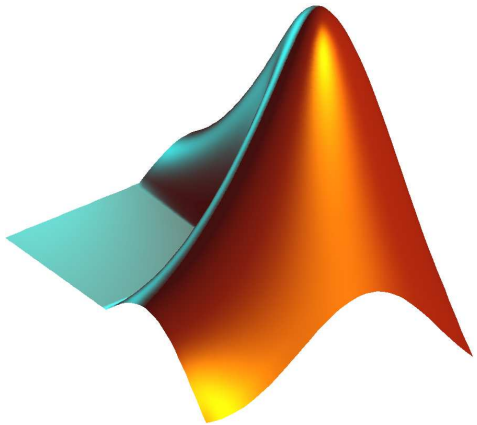
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



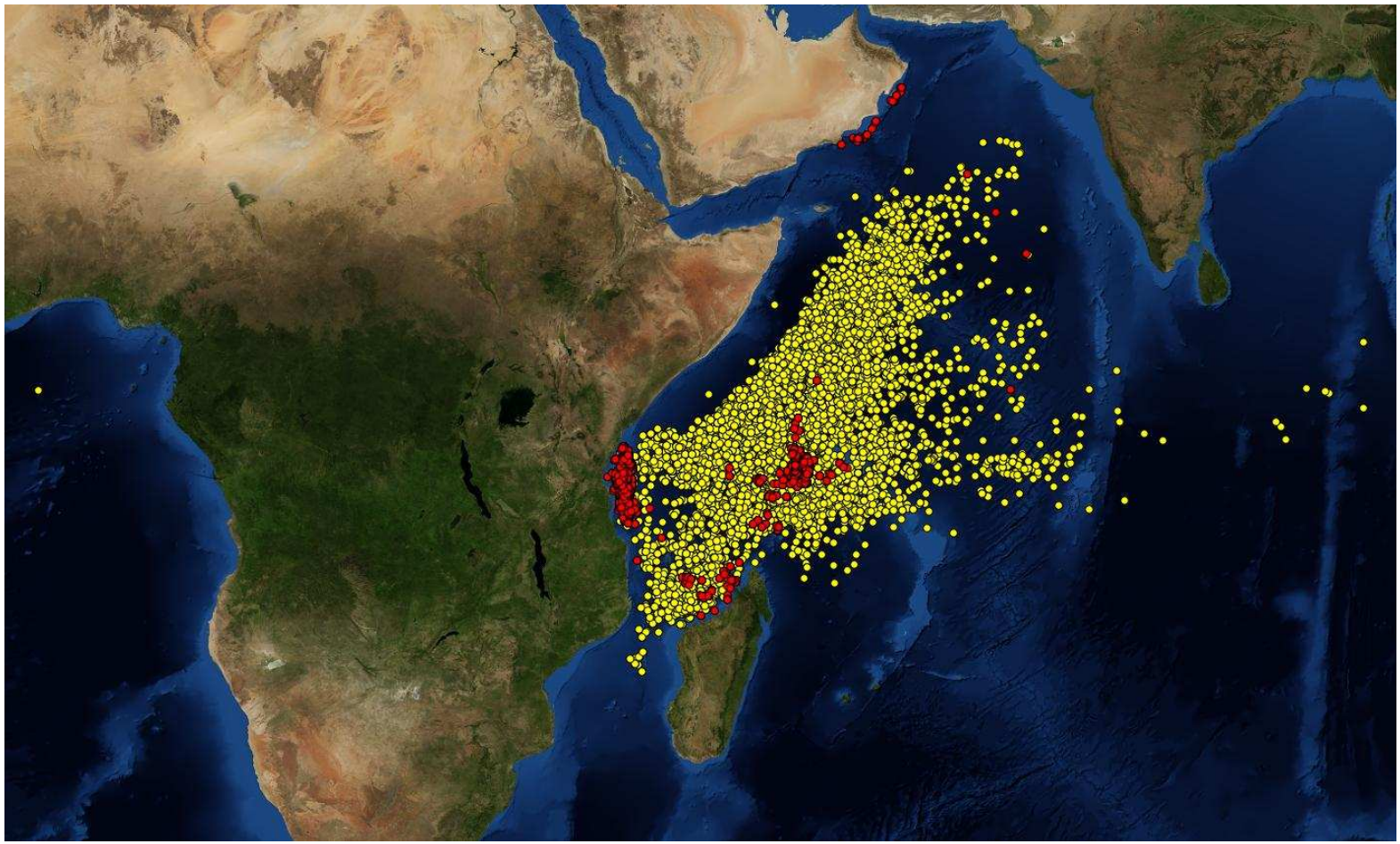
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



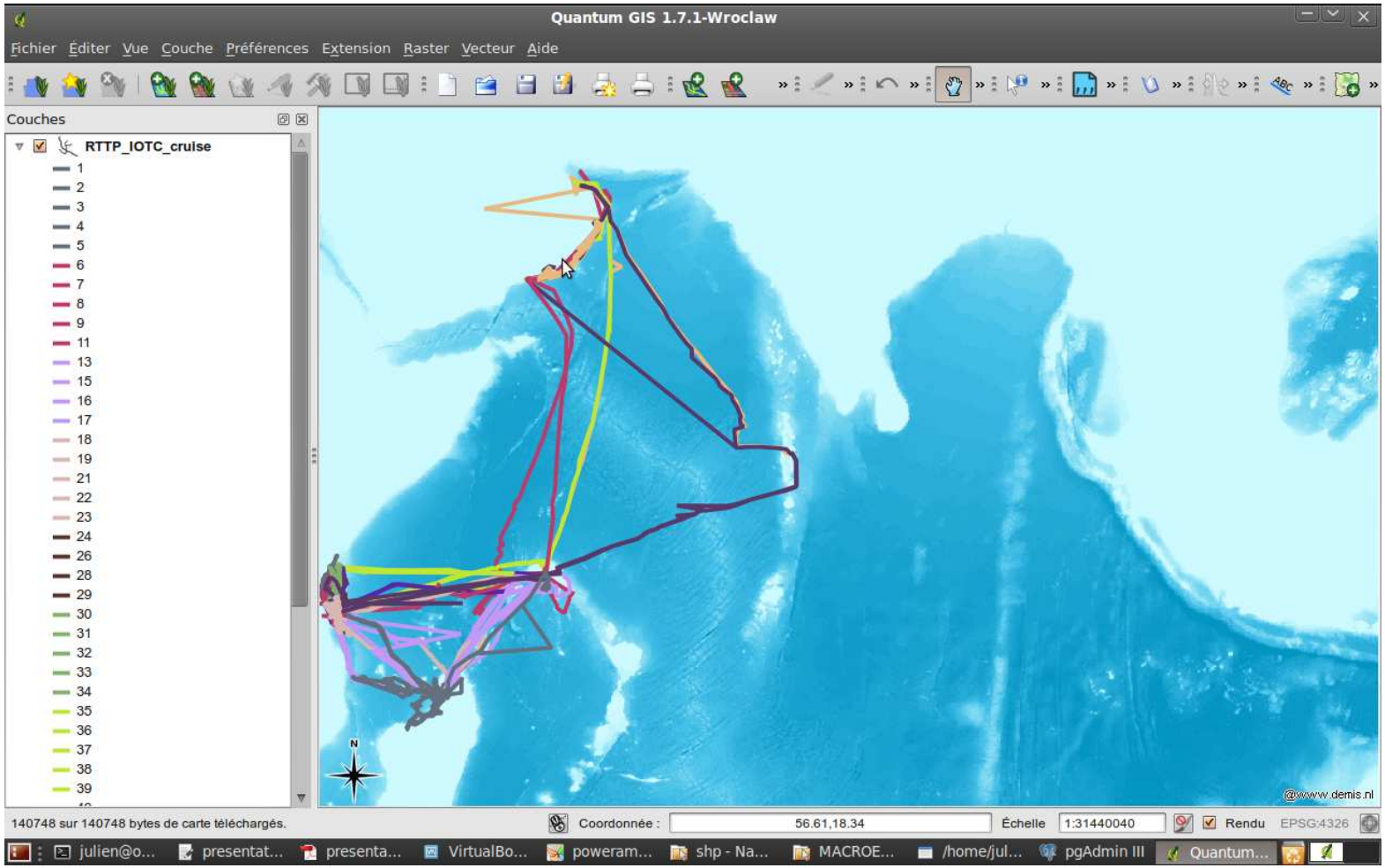
Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



Connection of data: hybrid approach of MDST (step 2 bis)

Similar GUIs wheter dataset has been transformed or not:



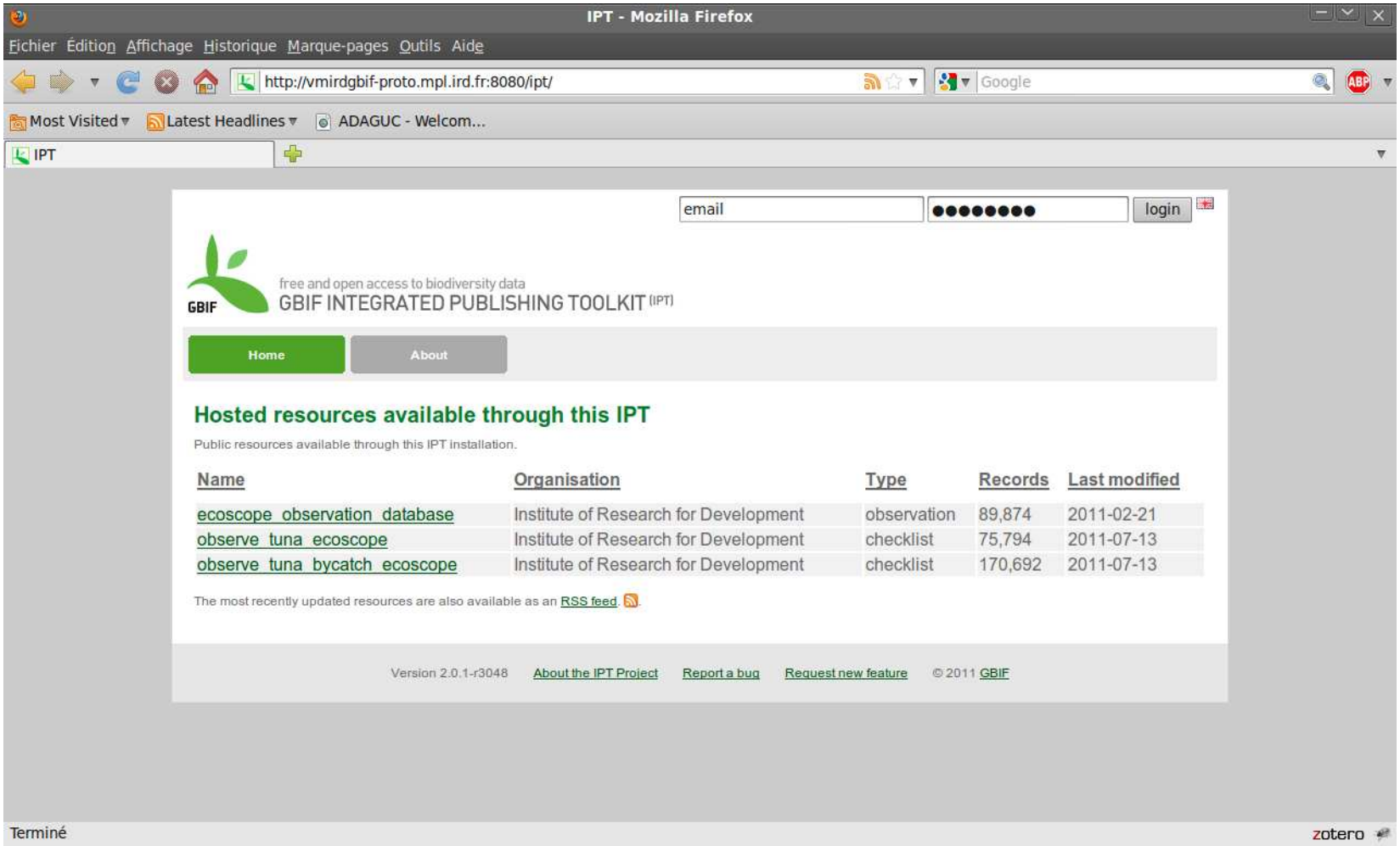
Data interoperability (step 3)

Data in this database are served with different standards according to the communities of users:

- Fisheries: Fishframe (for EU member states: DCF),
- Biodiversity: TDWG / GBIF,
- Spatial information with OGC standards (INSPIRE in Europe):
  - spatial databases enable to manage more than points: polygons, polylines... (real need for tagging data),
  - Postgis 2.0 enables to query usual RDMS data with environmental parameters managed in raster data formats (netCDF, geoTiff...).
  - same operations are possible with Web Services.
- Statistical data: SDMX,
- Web / Linked Open data: W3C,
- ...and others to come.

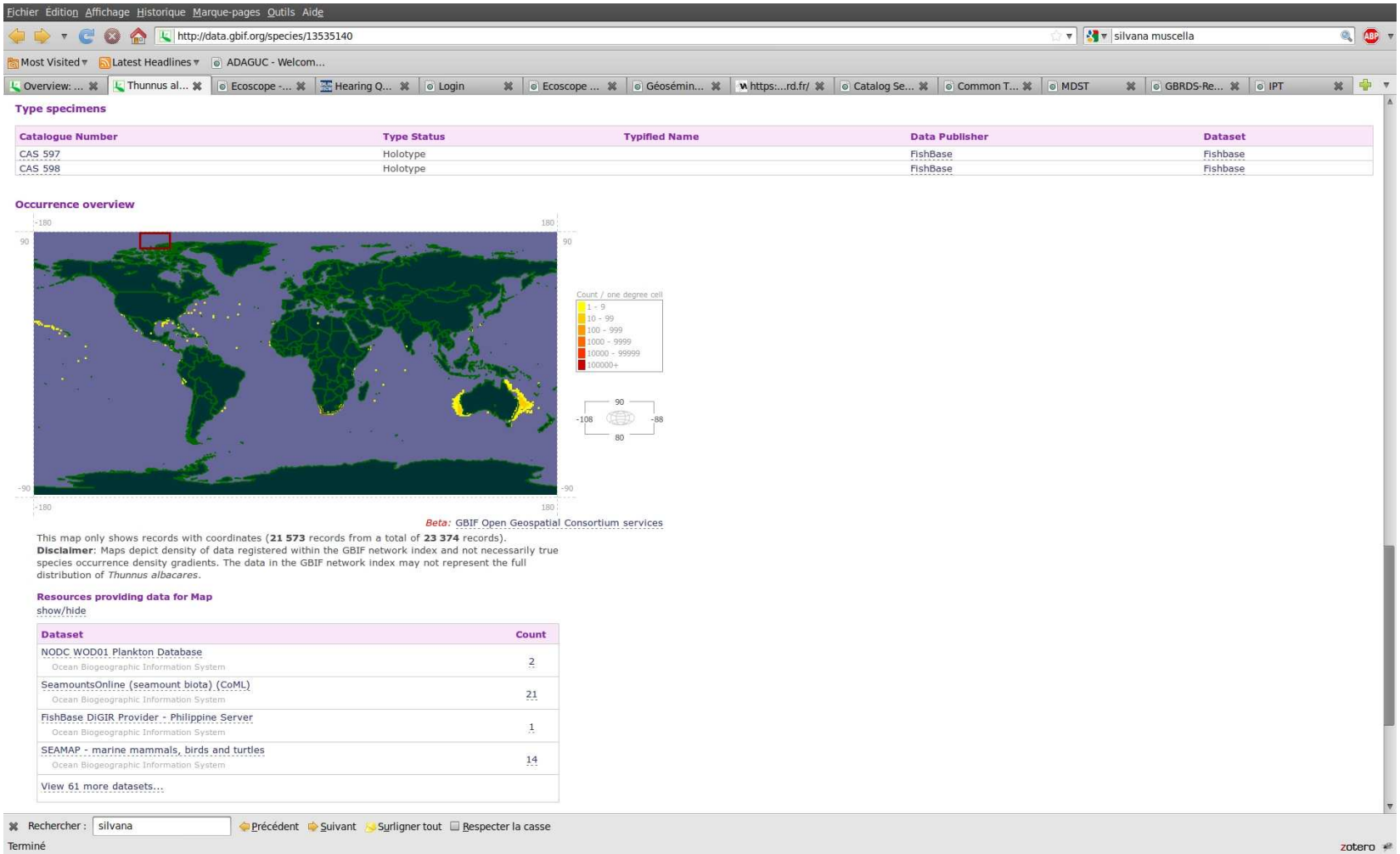
Interoperability results (step 3) for the generic database

GBIF (IPT server of IRD in Montpellier)



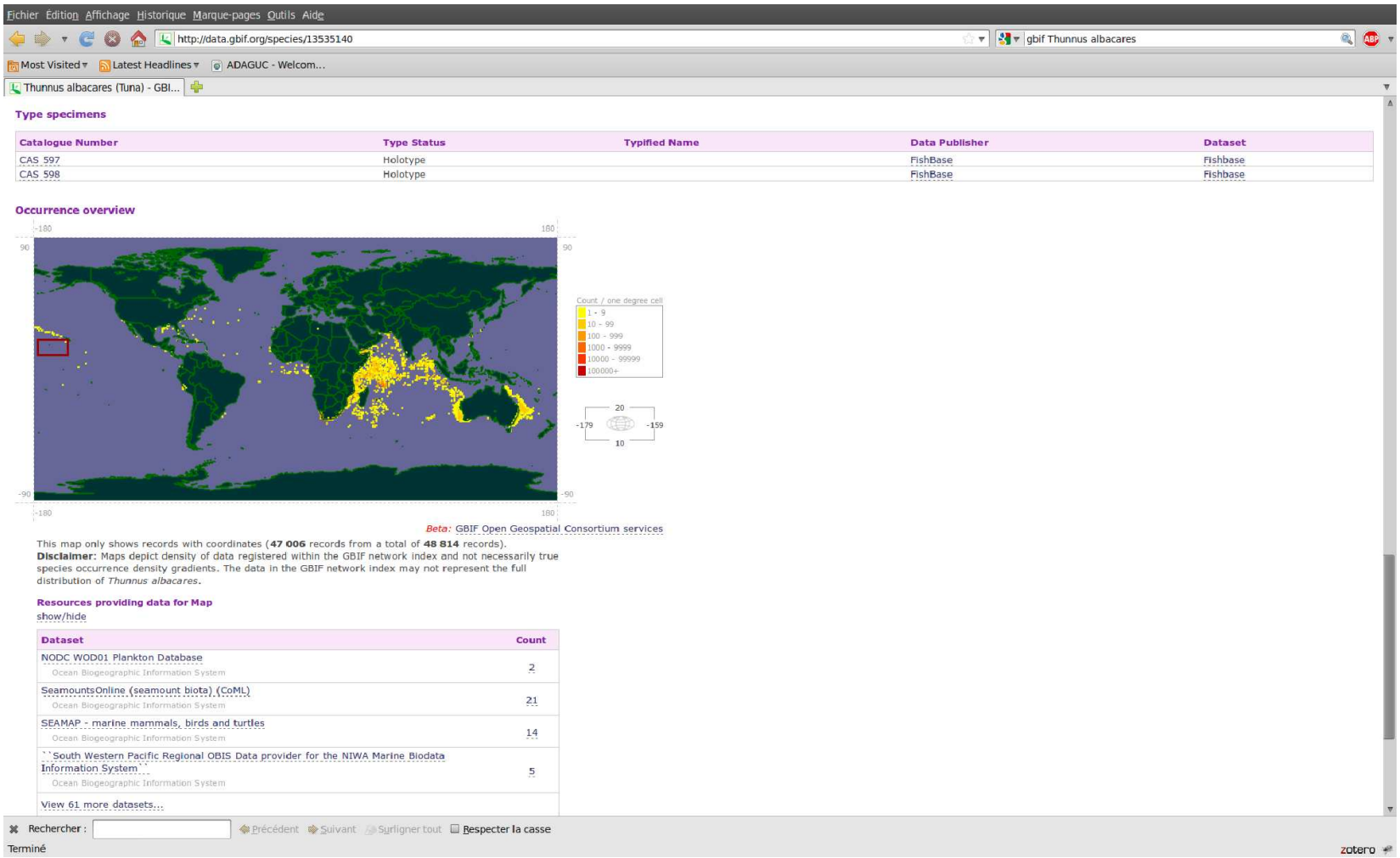
Interoperability results (step 3) for the generic database

GBIF (IPT server of IRD in Montpellier)



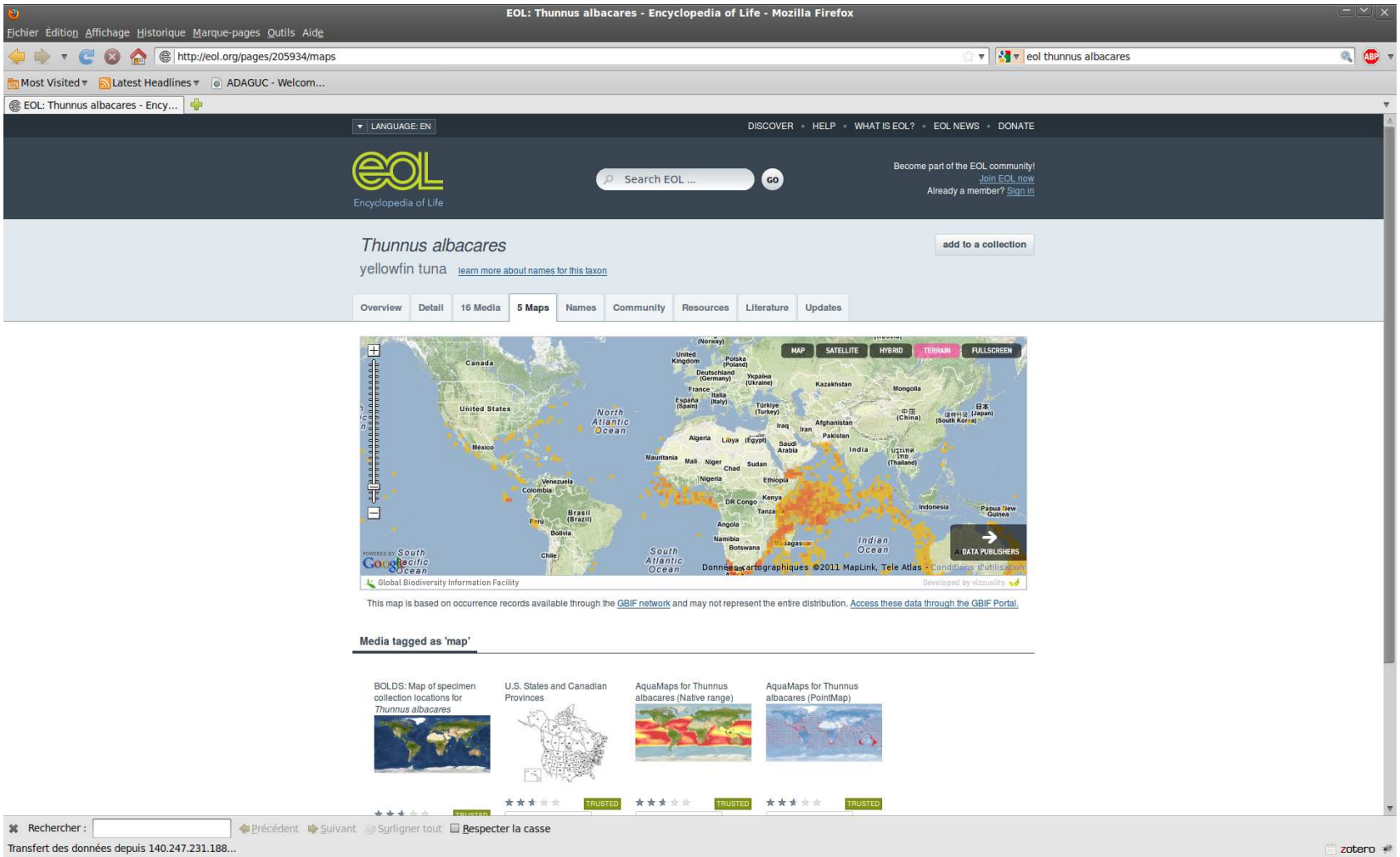
Interoperability results (step 3) for the generic database

GBIF (IPT server of IRD in Montpellier)



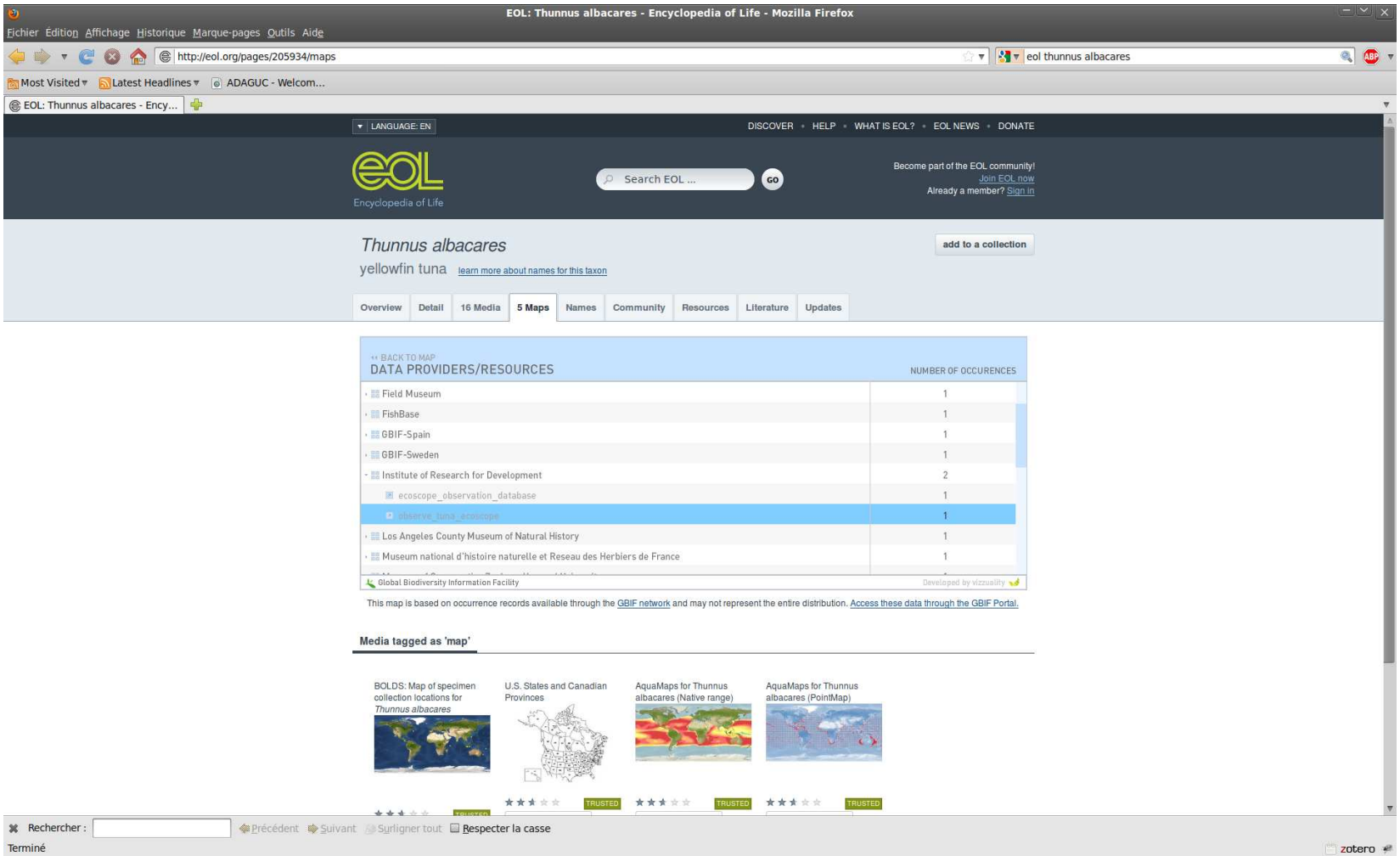
Interoperability results (step 3) for the generic database

Web portals via GBIF (EOL, OBIS. . .)



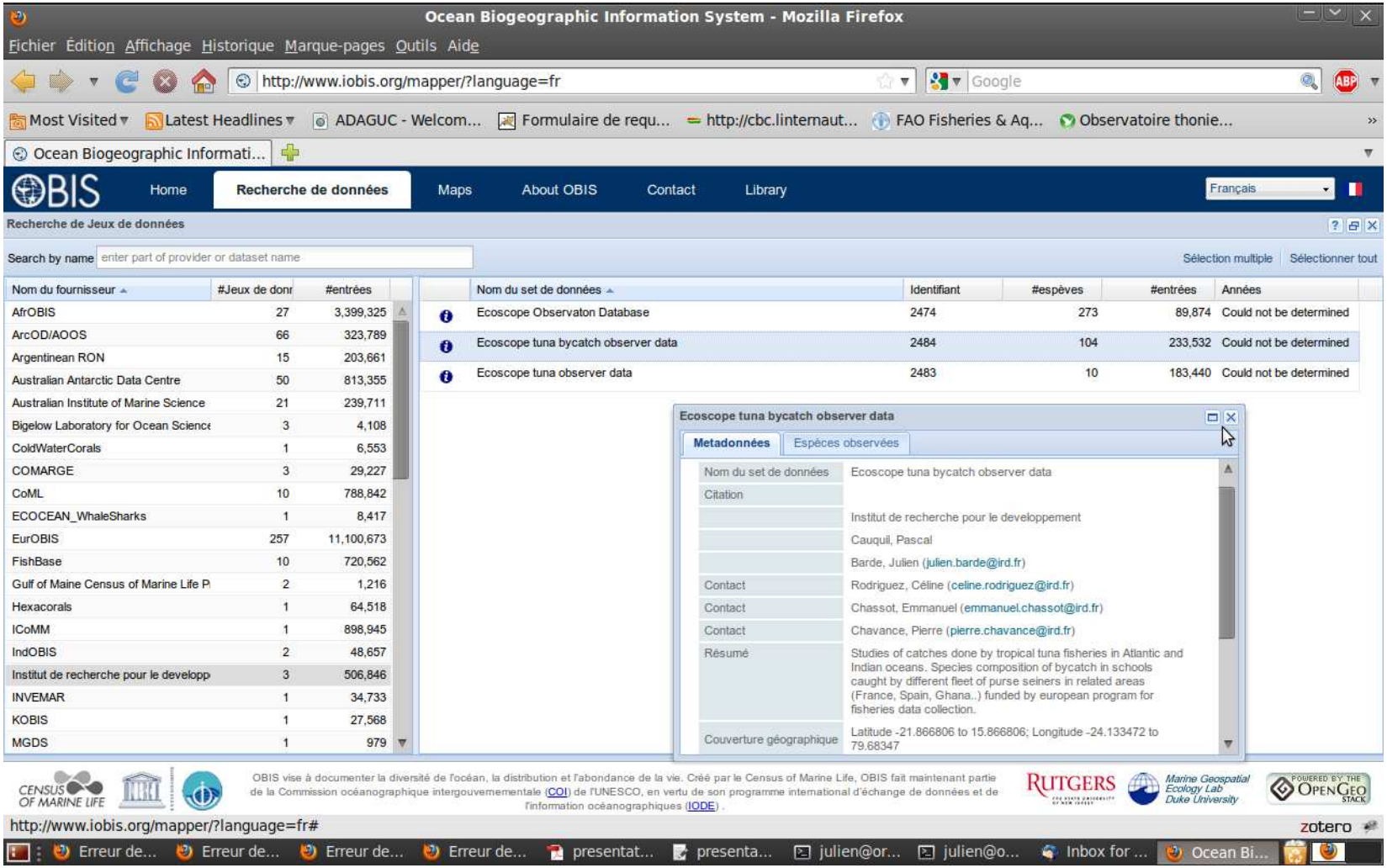
Interoperability results (step 3) for the generic database

Web portals via GBIF (EOL, OBIS. . .)



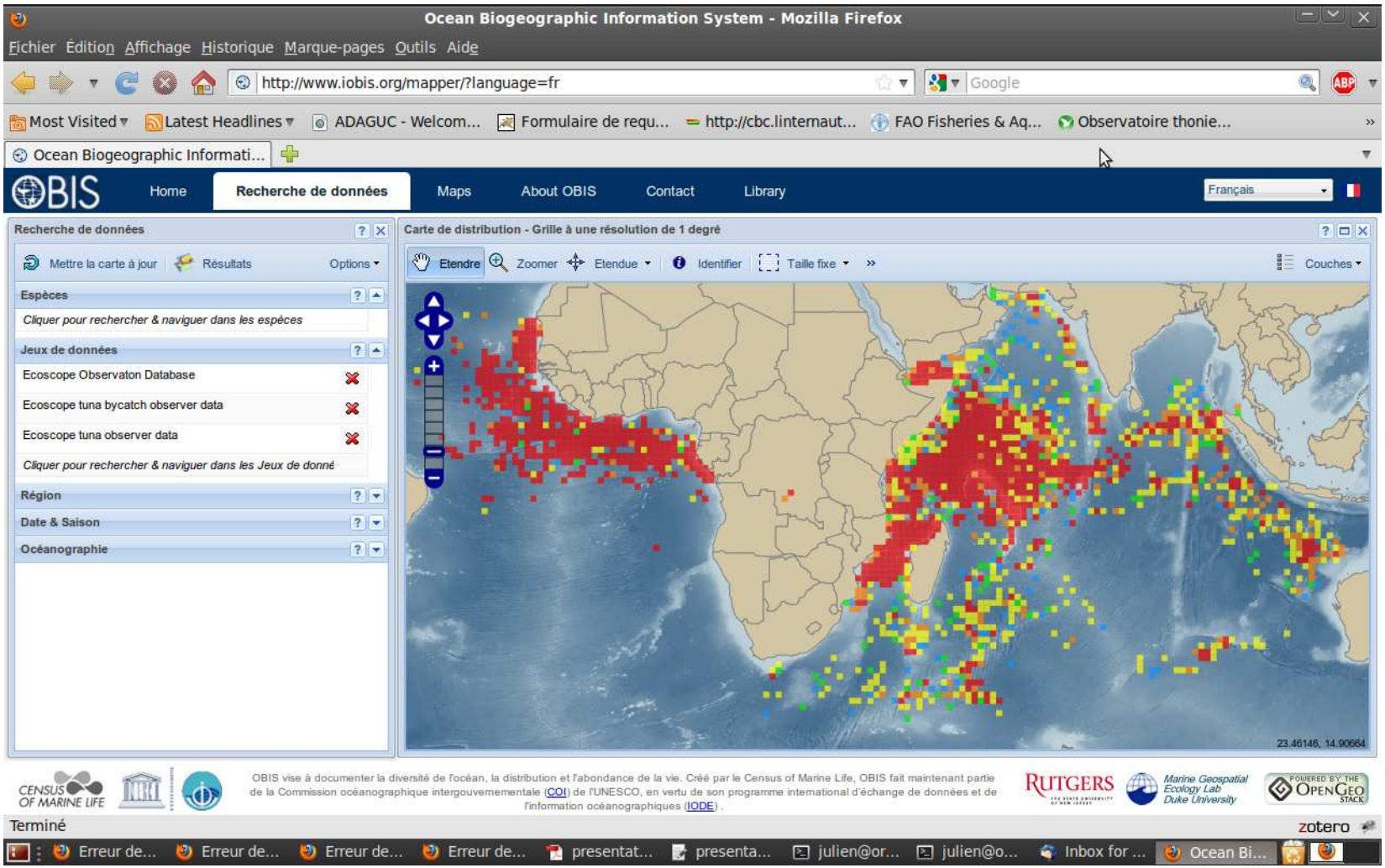
Interoperability results (step 3) for the generic database

Web portals via GBIF (EOL, OBIS. . .)



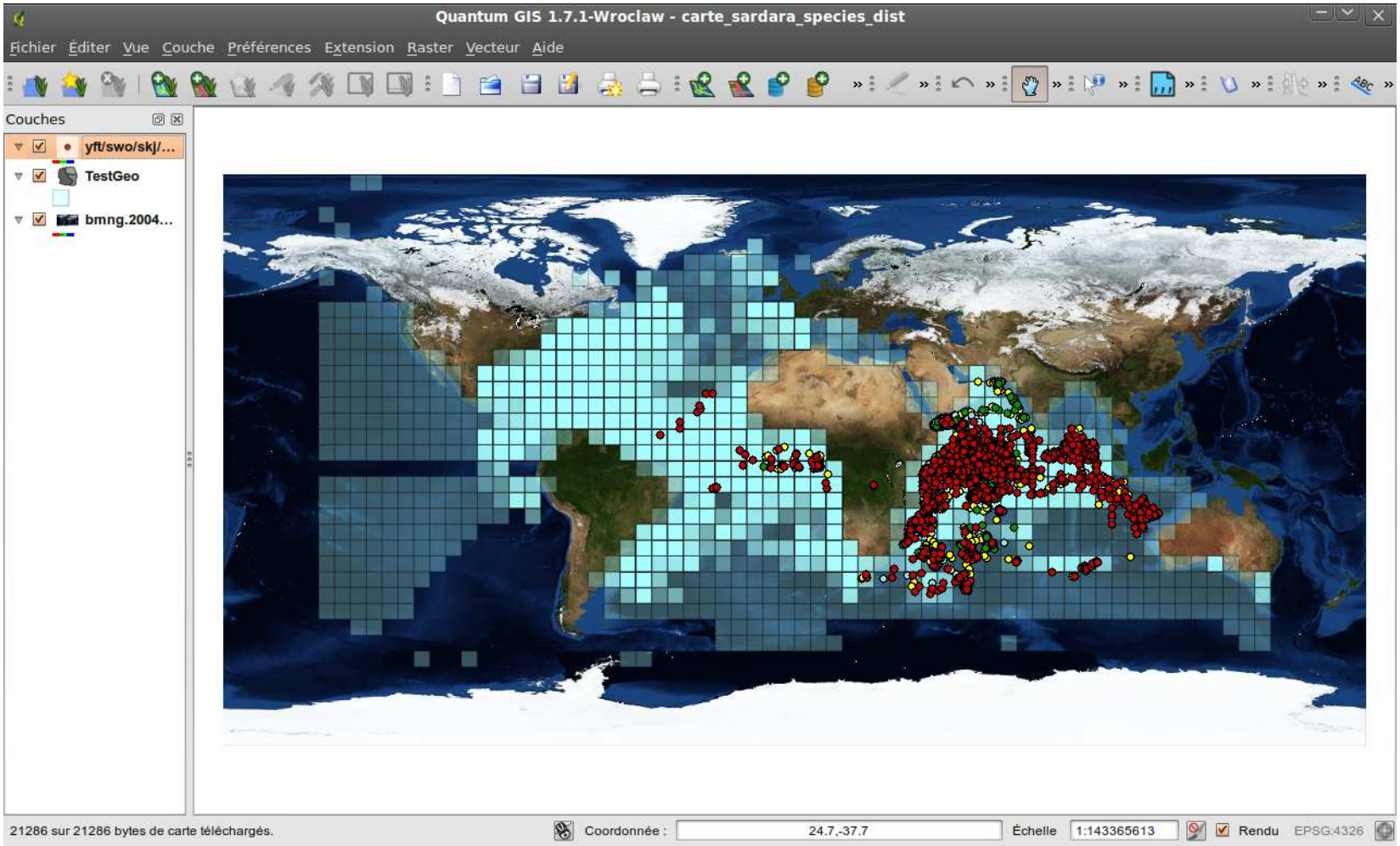
Interoperability results (step 3) for the generic database

Web portals via GBIF (EOL, OBIS. . .)



Interoperability results (step 3) for the generic database

GIS (via OGC standards / INSPIRE): Qgis, GvSig through Web Services



Summary

Achievements

Work done so far:

- new version of the database model to manage replicated observations,
- first steps of data migration and access OK:
  - tagging data and EAF data can be stored together,
  - data can be made available in different ways with a Web portal (MDST) through standards for data formats and access protocols,
  - can be blended with environmental data (Postgis + OGC).
- data can be connected to other infrastructures (GBIF, INSPIRE, ...) through standards / interoperability.

Next steps

Still a lot of work:

- to improve the model of the database,
- to separate raw data from estimated data . . .
- . . . by embedding libraries of processing in the database:
  - flag the quality of data,
  - location of the recovery with set of points (min, max. . .),
  - time at liberty. . . ,
  - density. . . ,
- loading of similar datasets: MAC (atlantic). . . ,
- administration interfaces for the database,
- publication of RTTP CTOI data on INSPIRE, GBIF, OBIS...with MDST ? Not plugged for now.

Additional information