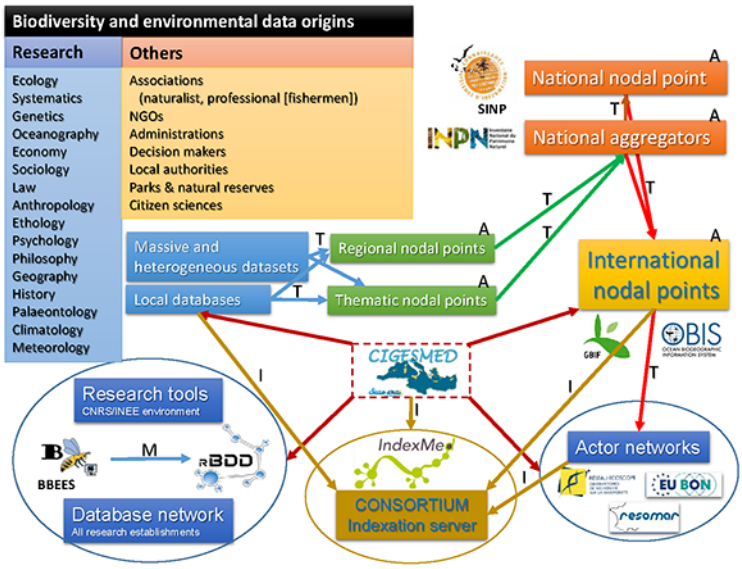
Ecological Data Preservation in the context of IndexMed - PREDON 2015

Romain DAVID – Jean-Pierre FÉRAL – Cyrille BLANPAIN
Corresponding author: romain.david@imbe.fr

Abstract

Ecology is still a poor sector of public investment. However, issues related to global change are largely measured through biological monitoring, and most human resources depend on these systems of human influence. Ecological research faces challenges to explain local, regional and global interactions in different human and natural contexts. To achieve these objectives, the "IndexMed" project plans to develop models and tools to allow data from different disciplines to be combined and to propose methods of scenarios creations by successive approaches, based on concepts currently described in the field of global ecology. The objective is to build graphs with heterogeneous data and their relationships concerning Mediterranean ecology and to analyze them in this new system.



Data and data management in global ecology: Synthetic scheme from the origin to the access and use of ecological data. All data are managed by the authors. There is no centralized database. Data are indexed, catalogued and made available thanks to special communication tools (plugins, fluxes).

Abbreviations.- **A**= data aggregation, **T**= data transfer, **I**= data indexation, **M**= management.
SINP: Information system on nature and landscapes, **INPN**: National inventory of the natural heritage, **BBEES**: Database on Biodiversity, Ecology, Environment and Societies, **rBDD**: Databases network, **GBIF**: Global Biodiversity Information Facility, **OBIS**: Ocean Biogeographic Information System, **CIGESMED**: Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters = example of an international program generating ecological data [www.cigesmed.eu].

Context and objectives of ecological databases in France

As specified in national and European laws and directives (INSPIRE, Aarhus convention…), ecological data in Europe must be accessible and free for use by the research community as well as other stakeholders. Scientifics research questions in ecology can be resolved at local, regional and global response scales by concomitant stakeholders only by combining data of different disciplines.

Data used by scientists in the field of ecology are no longer entirely produced by scientific institutions, which instead rely on external networks of participants and skills. At a regional scale, each factor can be measured by hundreds of different operators, organized in a territorial or thematic sense, but often disconnected from each other (between territories or thematic), and adopting protocols and data formats adapted to the problems of their duties (fishery, hunting, conservation, forestry, water management, research, agriculture, naturalist association, ...). However, there is no data producers' directory in the field of ecology.

For example, in August 2014, the data made available by the network supplying GBIF were made up of more than 500 million occurrences, which represent a small proportion of real world observations: at a regional scale, in France, occurrence data of Languedoc Roussillon represent less than 200 000 records in the international bank of GBIF, and in the same time, the data compiled by the regional information system reports 2 million records. Occurrence is the only one of hundreds or even thousands of possible parameters to be followed in multiple disciplines of ecology (landscape, ecosystems, taxonomy, epidemiology, ecotoxicology, human ecology, microbial ecology, molecular ecology, paleoecology, population ecology, restoration ecology, economic/social ecology, soil science, geography, philosophy ...).

Data aggregation systems are being developed at regional, national and international scales; they allow building large-scale reference information layers (Silenus PACA INPN, SINP, SISMER, ...). However, those will usually concern a data type or data collected in specific trades' goals, and no attempt has yet been made to create a system capable to make them usable together.

The development of participatory sciences leaves unanswered questions related to intellectual property and notions of responsibility with respect to this common heritage. The availability of these data is variable, and qualification processes that assess usability and efficiency are still rare.

No system permits a synthesis of existing data whatever the relevant competencies, discipline, type of factor or the data format, and creating necessary links between systems without trying to centralize all the data.

Conversely, production of scientific data is increasingly financed under the condition of accessibility (molecular biology data has been for decades, but for ecological and environmental data accessibility is only recently and rarely required). But the proper tools for this accessibility just don't exist…

IndexMed project: an original method to preserve and re-use ecological data at large scales

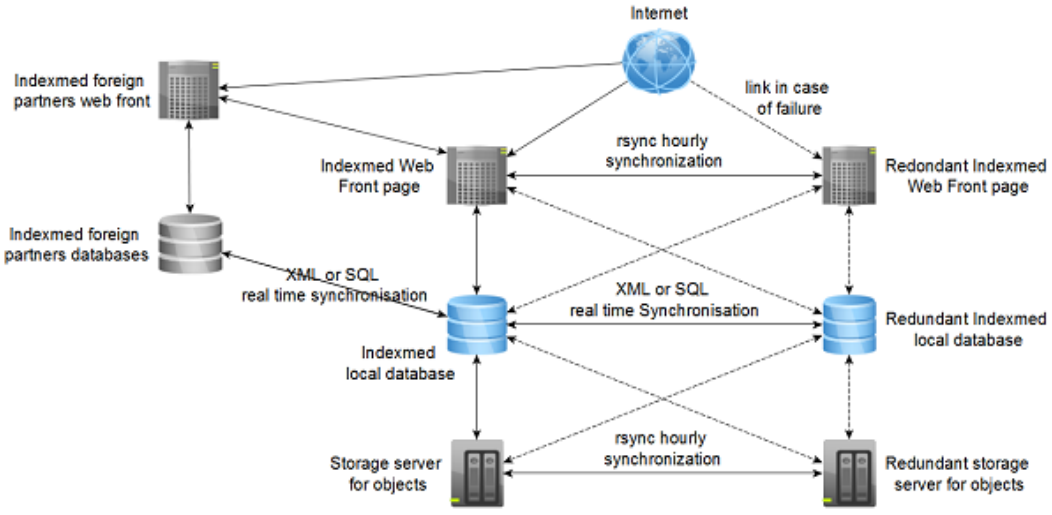
INTEROPERABILITY OF DATABASES IN ECOLOGY

Increased knowledge in ecology with a global approach requires comparing multiple data from different sources (physical data like temperature, depth, currents, biological data like occurrence, frequency, relative abundance, dominance, biomass taxa as well as economic or social data).

The main objective is to improve interoperability between different disciplines, and the “re-use” of data between all the users.
Considering that a database cannot contain all environment data and that a global ecological approach needs to compare data of several disciplines with lots of formats, standards evolving at their own pace, **building a centralized information system is simply impossible**.
The architecture of information systems for projects being developed must be decentralized and should consist of indexing, classifying, mapping, and interfacing of data from coastal and marine Mediterranean environments for research in ecology and management tools of natural spaces.
The proposed solution is to build solutions to link data and make them comparable **outside of each database**. For this, IndexMed project sets up a system of indexing data on Mediterranean ecology.

In each information center, data with an interest at a large scale will receive an identification number, or this data identifier will be linked to the indexing standard. For each modification of a record, the assignment of new identifiers associated with their history should enable version control and traceability needed to improve the quality and reliability of the data. The project will develop a resolution service that should value the results of ecological research by responding to current demands for social use (resolution refers to any process service that further identifies an object or entity from an associated, not-necessarily-unique alphanumeric name). It is based on large panels of participants, data and skills being developed.

DATABASES PRESERVATION and HARDWARE CONFIGURATION
This project requires a senior level Service to ensure data accessibility. For this we will create a redundant, synchronized computer architecture. The data that will be housed by the OSU Pythéas will come from observations of researchers or acquisition by sensors. These data will be hosted on storage servers that are redundant to ensure that data are not lost. Both storage servers are synchronized every hour with "rsync" and "rsnapshot" will possibly be added to make rollback possible. To reference this data we will use a database server, also redundant with real-time synchronization, which, depending on the choice made with our partners, will be based on the exchange of XML files or, using the native functionality, basic SQL-based to synchronize our local servers, but also to send the changes of our bases to our partners. Finally, users will have access to a reference web page where they can access public data locally hosted by the OSU Pythéas or definitions of data hosted by our partners. The stated goal is to ensure a 24 hour access emancipating us of any hardware failures.



Data and data Indexmed computing diagram

DATA QUALIFICATION

Linking data from several disciplines demands finding comparable properties between objects described by data. These properties may concern the availability, the quality of data as well as the property of objects (for example, a site can have properties about biomass level, number of species, human impacts, value of services provided by biodiversity and a species can be qualified by an international status of conservation, a life-history trait, ease of recognition…)

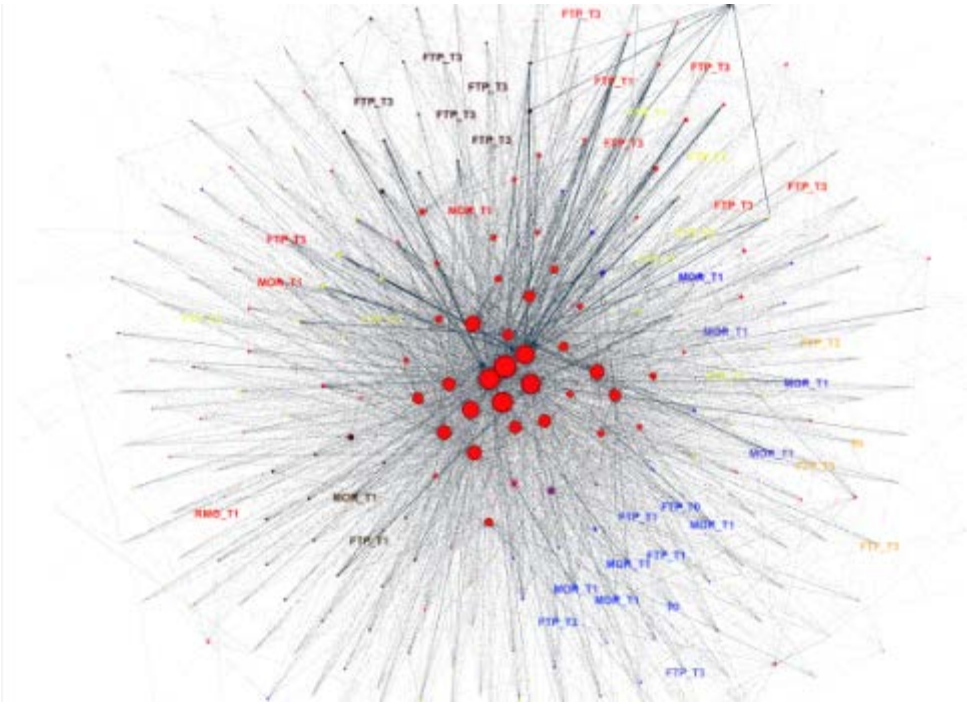
The IndexMed project is based on the possibility of re-qualifying or over-qualifying the index set up with different properties and match indexes of different formats and standards. All properties permit linking objects (species, observers, stations, indicators… concepts, conservation status …). Linking objects with different natures permit to construct “multi-object graphs”. This concept, inspired by those used by astronomers, requires participants to not only develop their metadata, but also to foster an increase for data *via* common "ontologies". One of the tasks of the future program will be to participate in the implementation of this ontology of Mediterranean biodiversity and propose tools and methods based on existing knowledge (Aix-Marseille University, University of Montpellier, CESAB ...).

NODAL POINTS

The result will permit installing nodal points of massive indexation and data classification of the coastal and marine Mediterranean environment. It will be followed by the development of configurable graphs to search in multidisciplinary data. A "modular" organization (for the administration of an object type or data independently by the most proficient actor) should be preferred rather than centralized systems, and mirroring systems will also be developed similar to domain name resolution systems.

To be re-usable by other communities and to demonstrate the power of exchange philosophy, the project will develop new transdisciplinary methods of data analysis, focusing on open data, open source and free methods and development tools. These indexing nodes will be "clonable" at will with enrichment rules and sharing relevant licenses "creative common" type "sharing the same conditions," allowing others to copy, distribute and modify the index provided and then disclose any adaptation of the index under the same conditions (open-source, open data). However, anyone who would like to publish an adjustment under other conditions must obtain prior authorization.

To implement its tools, the community includes today IndexMed skills and laboratories working in other fields than ecology, but encountering the same technology challenges. Arising from a common will, IndexMed is a technical challenge faced by many communities: IndexMed seeks the cooperation of mathematicians, physicists, computer engineers, archivists, astronomers, economists, and sociologists in order to best adapt the tools they use and rely on their expertise.



Example of graph in global ecology: Analysis of 350 photos distributed along 30 transects at 3 coralligenous sites (MOR: Morgiou, RMO: Riou-Moyade, FTF: Tiboulen du Frioul, Ti: transect number) by means of GEPHI v.0.8.2. It shows the species frequency for each transect, for different observers and methods. Differences between transects appear to be more important than between sites; and different observers do not introduce a huge variability.

IndexMed serves a scientific AND societal purpose. With an initial level of interoperability, the project will test the creation of "data cubes" by connecting for example, sociological, economic and ecological aspects of Mediterranean biodiversity. It will rely on the latest developments of IT and mathematics (algorithm graphs, data mining) to propose solutions to transdisciplinary issues that directly concern biodiversity. This consortium need more participants in each discipline and in more European countries to be efficient at the scientific and social levels: do not hesitate to propose a contribution or to join the involved teams (www.indexmed.eu)

Data cubes and graphs will permit combining different issues like for example factors relevant to describe good environmental status, values factors or combinations of factors relevant to measure a disturbance, or dynamic and precursors (associations between factor values) promoting the appearance of environmental perturbations (loss of efficiency, provision of ecological services ...). All these issues are important current social and scientific challenges.

Present and next step

IndexMed is a consortium created by the axis « Management of biodiversity and natural spaces » of the IMBE (Mediterranean Institute of Biodiversity and marine and terrestrial Ecology). This project was initiated as part of a response to the CNRS call for project MASTODONS (Large scientific datasets) and is now supported by the “Predon initiative”. Its main goal is to develop the knowledge of databases and their effective use in the ecological research community. This consortium responds to calls for projects using databases to address ecological issues in the Mediterranean basin, promoting multidisciplinary and collaboration with other entities of the CNRS and other European research centers. The projects to be developed by IndexMed members

must be based on various national and international initiatives and promote international collaboration. This consortium is particularly used as a bridge between existing networks and initiatives at national and international levels using data indexing and qualification.
A short-term goal of IndexMed is to establish a platform of data indexing of Mediterranean biodiversity and of environmental parameters which are of interest for research. This index will use the tools and methods recommended at both the national (SINP, MNHN / GBIF, RBDD) and international levels (MedOBIS, OBIS, GBIF, LifeWatch, GEOBON, …). It will be based on catalogs developed at national and later, international levels (IDCNP of the SINP, networks of actors of the FRB, and INSPIRE catalogs). Improved and completed, this project will be proposed at the call of project as a "work package" of a larger project on biodiversity: a middle term goal is to prepare a global and multidisciplinary project, supported by people from different research institutes, agencies and environmental societies, and thus increase the potential for future collaborations.

For more information:
Romain DAVID* romain.david@imbe.fr
Jean-Pierre FERAL* jean-pierre.feral@imbe.fr
Cyrille BLANPAIN** cyrille.blanpain@osupytheas.fr

*IMBE, Institut Méditerranéen de Biodiversité et d'Écologie marine et continentale (UMR 7263), CNRS/Aix Marseille University
Station marine d'Endoume
Chemin de la Batterie des Lions
13007 Marseille, France

** Observatoire des Sciences de l'Univers, Institut Pythéas
CNRS/Aix Marseille University
Pole de l'Etoile, Site de Château-Gombert
38 rue Frédéric Joliot-Curie
13388 Marseille cedex 13, France

Acknowledgment :Thanks are due to Abigail Cahill for improving the English text.

References:

Barde J., T. Libourel, P. Maurel. 2005. A Metadata Service for Integrated Management of Knowledge Related to Coastal Areas. Multimedia Tools and Applications, 25, (3): 419-429.
Peters, D. P. C., K. M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales. 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere 5(6):67. http://dx.doi.org/10.1890/ES13-00359.1
Robinson I., J. Webber, E. Eifrem. 2013. Graph Databases, O'Reilly Media, 224 pp. ISBN 13: 978-1-4493-5626-2, ISBN 10: 1-4493-5626-5