IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs

Romain David, Jean-Pierre Feral, Anne Archambeau, Nicolas Bailly, Cyrille Blanpain, Vincent Breton, Aurélie de Jode, Aurélie Delavaud, Alrick Dias, Sophie Gachet, et al.

International Environmental Modelling and Software Society (iEMSs)
8th International Congress on Environmental Modelling and Software
Toulouse, France, Sabine Sauvage, José-Miguel Sánchez-Pérez, Andrea Rizzoli (Eds.)
http://www.iemss.org/society/index.php/iemss-2016-proceedings

-

IndexMed projects: new tools using the CIGESMED DataBase on Coralligenous for indexing, visualizing and data mining based on graphs

Romain David[1], Jean-Pierre Féral[1], Anne-Sophie Archambeau[2], Nicolas Bailly[3], Cyrille Blanpain[4], Vincent Breton[5], Aurélien De Jode[1], Aurélie Delavaud[6], Alrick Dias[1], Sophie Gachet[1], Dorian Guillemain[1], Julien Lecubin[4], Geneviève Romier[5], Christian Surace[7], Laure Thierry de Ville d'Avray[1], Christos Arvanitidis[3], Anne Chenuil[1], Melih Ertan Çinar[8], Drosos Koutsoubas[3,9], Stéphane Sartoretto[10], Thierry Tatoni[1]

(1) Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE), CNRS, Aix Marseille Université, IRD, and Université d'Avignon, Station Marine d'Endoume, Chemin de la Batterie des Lions, 13007 Marseille, France. romain.david@imbe.fr, jean-pierre.feral@imbe.fr, anne.chenuil@imbe.fr, aurelien.dejode@imbe.fr, alrick.dias@imbe.fr, sophie.gachet@imbe.fr, dorian.guillemain@imbe.fr,laure.thierry@imbe.fr, thierry.tatoni@imbe.fr
(2) GBIF-France, MNHN, CP 48, 43 rue Buffon, 75005 Paris, France. archambeau@gbif.fr, gbif@gbif.fr
(3) HCMR/IMBBC Hellenic Centre for Marine Research, Institute of Marine Biology, Biotechnology & Aquaculture, LifeWatchGreece, Gouves, 71500 Heraklion, Crete, Greece. nbailly@hcmr.gr; arvanitidis@hcmr.gr
(4) Service informatique (SIP), OSU Pythéas, CNRS, Aix Marseille Université, 13007 Marseille, France. cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr
(5) Institut des Grilles et du cloud (IDG) France Grilles % LPC Clermont-Ferrand 4 avenue Blaise Pascal 63178 Aubière Cedex breton@idgrilles.fr, genevieve.romier@idgrilles.fr
(6) FRB ECOSCOPE - Pôle pour l'observation et la diffusion des données de recherche sur la biodiversité, Fondation pour la Recherche sur la Biodiversité, 195 rue Saint-Jacques - 75005 Paris, France. aurelie.delavaud@fondationbiodiversite.fr
(7) Laboratoire d'Astrophysique de Marseille (LAM), CNRS, Aix Marseille Université, rue Frédéric Joliot-Curie, 13013 Marseille, France. christian.surace@lam.fr
(8) Department of Hydrobiology, Faculty of Fisheries, Ege University, Bornova, Izmir, Turkey. melih.cinar@ege.edu.tr
(9) National Marine Park of Zakynthos, 29100, Zakynthos, and Dept. Marine Sciences, University of the Aegean, 81100 Mytilini, Greece drosos@aegean.gr
(10) IFREMER,Quartier Brégaillon, 83500 La Seyne-sur-Mer, France. Stephane.Sartoretto@ifremer.fr

Abstract: Data produced by the *SeasEra* CIGESMED project (Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters) have a high potential to be used by several stakeholders involved in environmental management. A new consortium called IndexMed whose task is to index Mediterranean biodiversity data, makes it possible to build graphs in order to analyse the CIGESMED data and develop new ways for data mining of coralligenous data. This paper presents the prototypes under development that test the ability of graphs dataBases and tools to connect biodiversity objects with non-centralized data. This project explores the ability of two scientific communities to work together. The uses of data from coralligenous habitat demonstrate the prototype functionalities and introduce new perspectives to analyse environmental and societal responses.

Keywords*: data qualification, graph, thesaurus, distributed information system, coralligenous habitats*

1 INTRODUCTION

1.1 Context: big data and interoperability in ecology

Data mining emerged in the late 90s [Fayyad et al., 1996] as a discipline to extract relevant novel and understandable knowledge from the analysis of pre-existing datasets and evolved to an increasingly complex approach which includes ecology, among other disciplines. Although currently it is considered,

by most information producers and users in scientific disciplines and industry, as the most promising way for making progress and leading to discovery, the use of big data in ecology is still lagging behind other disciplines [Peters et al., 2014]. Regarding marine biodiversity and its connection with the coastal socio-ecological systems (SES), data production is still very expensive and with a low level of automation. Studies on long term data series and/or large spatial areas are difficult to conduct, and when several observers need to be involved, the robustness and reproducibility of the observation is very often more difficult to obtain.
In a framework production multi-source data in ecology, the equivalence of observation systems and inter-calibration become crucial. Increasingly, integrative transdisciplinary approaches become necessary in the study of systems where information in each discipline is patchy, imprecise and poorly distributed. Yet all variables (biotic, abiotic, anthropogenic and natural pressures, perceived and recognized ecosystem services, societal perception, etc.) of these systems interact in a wide range of spatio-temporal scales [Féral et al., 2014] [Gachet et al., 2005], [Conruyt et al., 2010]. Some research systems tried to bring out logical interdependencies in socio-ecological systems to facilitate the building of biodiversity and ecosystems services [Laporte et al., 2014]. Several authors and international initiatives also tried to specify, through a hierarchical approach of biodiversity [Noss, 1990], a common minimum set of variables to be measured, complementary to one another and covering the interlinked biodiversity organization levels. They should allow to capture, with current means and tools, the maximum possible information on biodiversity state and trends with the least effort [Pereira et al., 2013], [Kissling et al., 2015]. Similar initiatives are ongoing for climate, weather and ocean [Connecting GEO] to foster the discovery and the analysis of complementary data across spatial and temporal scales.
New opportunities are created by open data formats in ecology [Reichman etal., 2011] and qualification standards usable in data management are developed with the Biodiversity Informatics Standards (formerly Taxonomic Database Working Group) consortium <www.tdwg.org> (Darwin Core Task Group) [Wieczorek et al., 2012]. Other studies focus on the integration of declarative knowledge with numerical and qualitative data [Gibert et al., 2014] or on the post-process of results required to provide understandable knowledge to the end-user [Cortez et al., 2012] [Gibert et al., 2012].

Finally, methods for linking biodiversity and environmental data exist, but they are often limited to an "inventory" aspect of biodiversity (collection, observations, repositories and distribution) and neglect functional aspects. Initiatives like CoL [Catalogue of Life], Data-ONE [Data Observation Network for Earth], EMODnet [European Marine Observation and Data Network], GEO-BON [Group On Earth Observations Biodiversity Observation Network], EU-BON [European Biodiversity Observation Network], GBIF [Global Biodiversity Information Facilities], LifeWatch, OBIS [Ocean Biogeographic Information System], and TDWG [Biodiversity Information Standards] along with Darwin Core and ABCD [Access to Biological Collections Data], are well-known examples for achieving interoperability and standardizing data collection. However, integrative approaches in the coastal management zone need more interoperability at each scale [Féral and David, 2013].

1.2 Coralligenous habitat's case

The "coralligenous habitat", an endemic bioherm of the Mediterranean Sea, offers such a particularly complex case of data management. Coralligenous habitats are difficult to study because they are patchy, not easily accessible (between 20 m and 120 m deep) and highly variable in local contexts [Ballesteros, 2006]. Due to these difficulties and the intrinsic complexity of this habitat type, comprehensive studies were rare until the 2000s [Laborel, 1961], [Laubier, 1966], [Hong, 1982], [Sartoretto, 1994]. Most of the proposed monitoring protocols / indicators for its ecological health are developed locally or regionally [Deter et al., 2012], [Sartoretto et al., 2016], on a single type of this habitat [Pergent-Martini et al., 2014], [Sini et al. 2015] and use rapid assessment techniques [Bianchi et al., 2007], [Kipson, 2011], [Deter et al., 2012], [Gatti et al., 2015], depending on prevailing environmental conditions.

Coralligenous habitats have been systematically studied at a larger scale within the CIGESMED ERANET'S program (*Coralligenous based Indicators to evaluate and monitor the "Good Environmental Status" of the MEDiterranean coastal waters*). The main CIGESMED's goal was to understand the connections between pressures (natural or anthropogenic) and the ecosystem functioning in order to define and maintain the Good Environmental Status (GES) of the Mediterranean Sea, by studying the typical, complex and poorly known habitats built by calcareous encrusting algae: the coralligenous habitats. This program is in support of the implementation of the Directive 2008/56/EC of the European Parliament and of the Council of the 17th June 2008. It participates establishing a framework for

stakeholders community action in the field of marine environmental policy (Marine Strategy Framework Directive - MSFD), and highlighting descriptors 1 (biological diversity), 2 (non-indigenous species) and 6 (seafloor integrity). The Marine Strategy Framework Directive (MSFD) is directing European Member States towards an implementation of the assessment of marine environmental status. Due to their very high specific richness, including commercial species, and the number of aesthetically important seascapes they hold, coralligenous habitats are one of the most popular marine environments [Ballesteros, 2006]. The community of CIGESMED redefined it as: "*reefs in dim-light conditions mainly bio-constructed on hard substratum by calcifying coralline algae widespread throughout the Mediterranean Sea, including patchwork of habitats complicated by the action of bio-eroders. These complex biogenic formations provide a number of different conditions of light, food and shelter. They are often considered as biodiversity hotspots gathering numerous sessile and sedentary species such as sponges, bryozoans, corals and gorgonians depending on the region and on the depth, to which hundreds of sciaphilic species are associated. These complex environments are a reservoir of natural resources (fisheries, red coral) and form highly valued landscapes sought by divers*". The data provided by CIGESMED are now used by the IndexMed consortium as a model, for developing data mining and decision support.

1.3 IndexMed, an open consortium

IndexMed is a new consortium in charge of indexing Mediterranean biodiversity data, building and analyzing graphs from heterogeneous databases. It aims to develop new ways for data mining in ecology. This consortium aims to identify and overcome the scientific barriers encountered when working on the quality and heterogeneity of data. The use of emerging data mining methods like graph-based models and analyses allows us to address these issues for improving decision-making support. These methods enable us to detect new patterns of contexts factors, invisible when using multidimensional analyses that have an accurate capacity to indicate particular situations [Klemes, 2015].

2 DATA INTEGRATOR SYSTEM

2.1 The challenge of quality data management to enhance results of data mining

Besides theoretical scientific issues (such as the intrinsic heterogeneity and complexity of biodiversity data, from genes to ecosystems, and their links to environmental parameters), the improvement of data quality is hindered by data management issues, such as: i) the dynamic update of voluminous datasets, ii) the update of reference repositories and standards supporting data management, iii) the heterogeneity of data producers and their motivation to maintain and supply their information systems, and iv) the diversity of the targeted end-users and their skills.

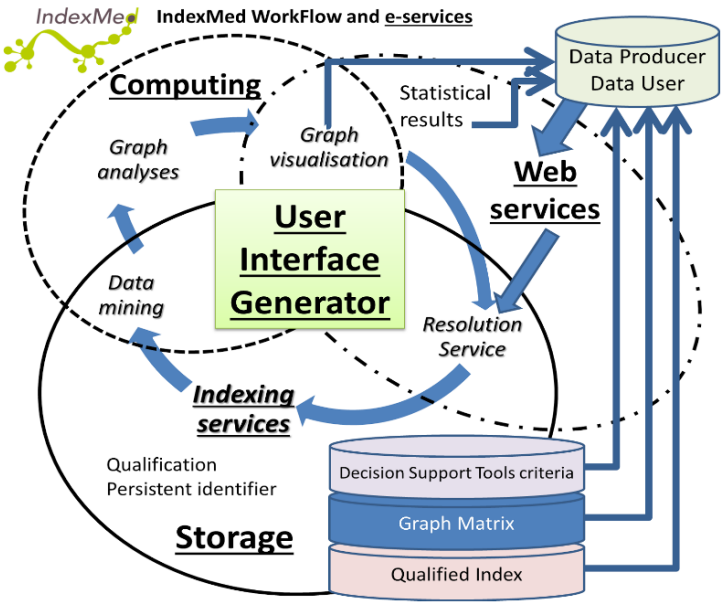
An integrated approach was implemented to mutualize heterogeneous dataset from different scales and disciplinary fields, to allow visualizing large data collections, and to extract new multidisciplinary knowledge in order to study complexity of coralligenous ecosystems. Health quality Indicators, targeting different levels of biodiversity (from communities to genomes), were co-constructed and tested by scientists, stakeholders, and by a citizen science network. Within the CIGESMED program, an upgrade of the design of each protocol and the inter-calibration exercises between various observers, materials, methods and organizations allowed to obtain: i) an assessment of the data variability due to natural or anthropogenic conditions and ii) a comparison of the different methods or observers and their efficiency. It showed that, under the above consensual definition of coralligenous, coralligenous habitats are made of a large panel of different species assemblages. For instance, coralligenous habitats from the Eastern and Western Mediterranean basins may share only 2 or 3 conspicuous species. Comparisons between regions are complicated by this lack of common species and the environmental conditions that can deeply change. Other links between typical environments and species (e.g. traits, contexts, structures, etc.) should be used to build indicators and to compare the environmental status between regions at a Mediterranean scale. Only multi-criteria contextualization by common factor value level allows constructing and adapting the indicators at a local scale and highlighting the significance of this indicator. Finally, it was decided to use competencies and tools developed by the IndexMed consortium to analyse all heterogeneous data, and integrate multidisciplinary data related to coralligenous habitats within the same multi-criterial approach but considering them at a comparable level of importance.

Coralligenous habitats are under anthropogenic uses and threats [Ballesteros, 2006]. Elaborating and testing “co-interpretation methods” of analyses (i.e. different datasets from different disciplines in the same time, at a same level of integration) is thought to be a keystone to mix in the same studies these heterogeneous data from different socio-economics disciplines and biodiversity disciplines, from the genome to the seascapes level. The ultimate goal is to propose scenarios to reach the sustainable management of biodiversity balancing exploitation and conservation. Data mining methods will be able to bring new perspectives to the disciplinary researches that finally examine interrelated objects (e.g. environmental chemistry, genomics, transcriptomics, metabolomics, population ecology / landscape, socio-ecological systems).

2.2 Workflow and e-services

To be able to use different and distributed datasets for data mining, a prototype of “object resolution service” (i.e. a web service that finds links and dependencies among indexed objects, based on unique objects identification (Figure 1)) that can be replicated by stakeholders is shared on a nodal point. It is intended to be replicable as a free software and a free service from European grids (EGI and others).
The aims of this prototype are to i) list available data and data stream, ii) analyse content of datasets and data streams with standards referential, iii) qualify streams, datasets with unique identifiers if there is no identifiers, iv) suggest matches between fields to users /matches between equivalent data rows. The role of this object resolution service is to establish links between data row with different “unique identifiers” (e.g. different versions of data raw, interdependencies between raw data and transformed data, etc.).

Figure 1 - IndexMed Workflow and e-services:



The resolution service is able to compare the index with storage data in e-infrastructures and other distant XML, JSON Flux from different databases. When necessary/possible, it creates a persistent identifier or links datasets or data records with existing identifiers. A scientist interface, adapted to the level and needs of each user allows a qualification process. The indexing service accept/manages data for computing services like data mining and graph analyses, and statistical results and graph models are stored and proposed as a new persistent flux.

When it is possible, data qualification uses tools, standards and recommendations at both national (SINP [National Information System on Biodiversity], RBDD [Network of Research Databases]) and international levels (MedOBIS [Mediterranean Ocean Biogeographic Information System], OBIS, GBIF [Cryer et al., 2009], Life-Watch, GEO-BON, etc.) or shared by other research entities (i.e. IRD [Institute of Research for the Development] or MNHN [National Museum of Natural History, Paris]).
Heterogeneity in datasets may be the result of a lack of standards to name and describe data [Kattge et al., 2011; [Madin et al., 2008]. Thus, attention must be paid to the characterisation of concepts by using “controlled vocabulary” (i.e. with a shared definition commonly choose) and semantic links between these concepts, which implies building a thesaurus in the first place. Thesaurus appears more appropriate than ontology because of its flexibility. Several eco-informatics initiatives attempted to build such thesaurus (see [Michener & Jones, 2012; Laporte, 2012]) and it is expected to take them in account.

New data qualifications generated by IndexMed prototypes aim at following the “guidelines on Data Management in Horizon 2020” (V2.1, 15 Feb. 2016) and cover, as recommended, the handling of research data during and after the project : data type, data collection (processed or generated), methodology of collect, standards used, post project data availability and access (open access as it is

possible), data maintenance and curation (Horizon 2020 Annotated Grant Agreement for articles 29.2 and 29.3, IP/12/790 on open access in Horizon 2020. <http://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/open-access_en.htm>)

3 CASE STUDY

The first tool, developed commonly by the CIGESMED and the IndexMed communities for scientists working on biodiversity, is a prototype web interface building dynamic maps of data and their possible links based on the graphs theory [David et al., 2015]. It can be used to establish links between objects of different disciplines and is able to connect data without centralizing them. The first aim is to teach the community on how to use the graph approach, featuring a didactic and ergonomic interface (Figure 2) with the aims to improve by step user level. It allows evaluating the data quality level and identifying the best ways to improve their efficiency (e.g. density, sensibility, velocity, accuracy, etc.). A tool permits to keep the new models designed by users (i.e. link and node selections) and new items (more than 1 object in a node) and produces a single stream usable by data centers at different formats (NoSQL exports in RDF, Json, XML formats) with a persistent URL.

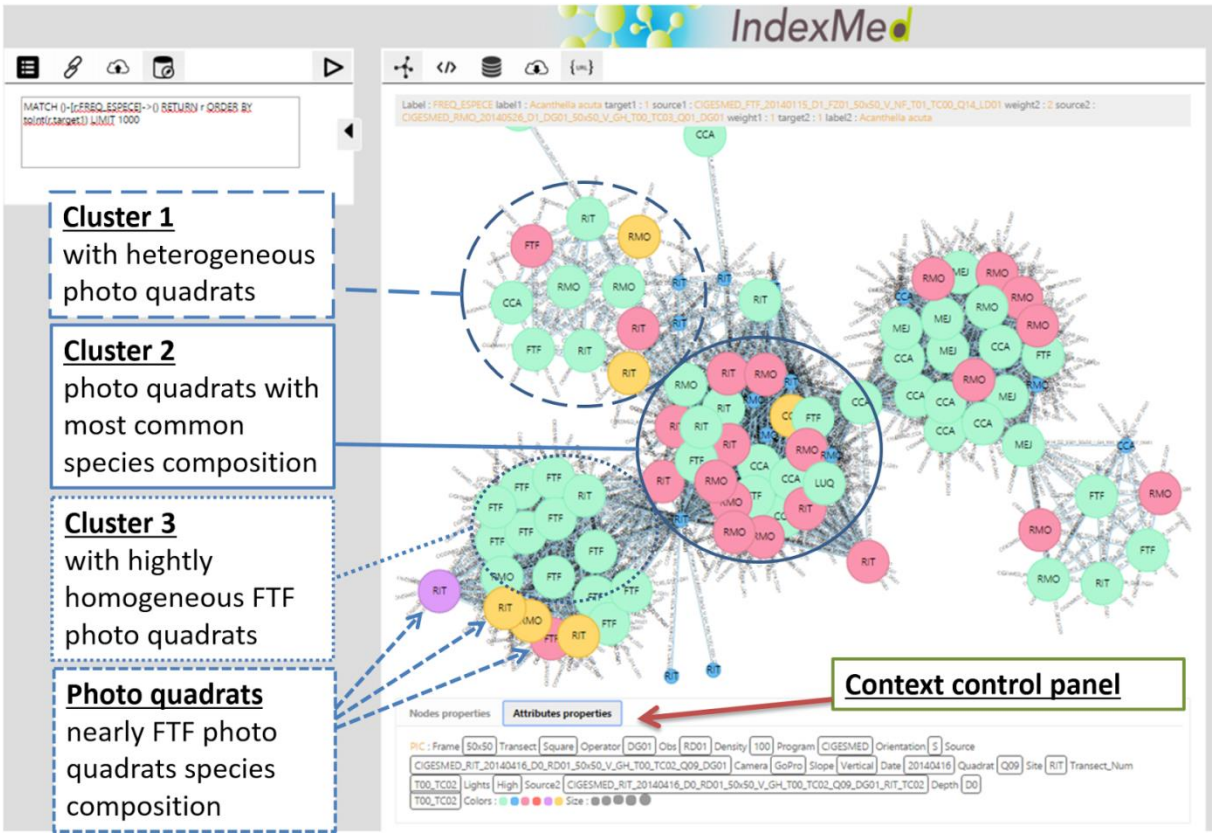


Figure 2: Example of an application of the prototype graph to a data set of 100 photo-quadrats made on coralligenous habitats in Marseille.

In the case presented in figure 2, photo quadrats selected by the interface are the nodes of the graph, species frequency selected by the user build 1000 links between photo quadrats, the colours of nodes highlight different elements of contexts (here, different observers). Node legends are the names of the observed sites. The more photo quadrats contain similar species assemblages, the more they are attracted. We can observe that some photo quadrats of different sites are attracted, thus similar in term of species frequencies, and very ubiquitous (cluster 2 in the centre of the graph). In cluster 3 photo quadrats are homogeneous and typical of the site. In the rim of cluster 3, some photo quadrats of other sites constitute a particular group. Cluster 1 shows another type of photo quadrats, less present in each site but represented everywhere.
In the example of figure 2, datasets come from 3 different protocols and data production systems, including one based on photo quadrats analyses with the software photoQuad [Trigonis and Sini, 2012],

a cartography of ecological/physical contexts and genetic data. Data objects represented on the graph are photos coming from different sites. Objects can be selected using the context control panel. Clusters visible in figure 2 can be modified by selection of context or species choice for the links. The relative importance of each species can be modified in the links (e.g. depending of the status or a specific trait) and some context elements can be selected to participate to links between nodes. Observers, highlighted by colours of nodes are not evenly distributed. Experience of observers is reflected in the size of nodes.

This interface uses indexed data, data qualification, and data traceability for discovering patterns in the conjunctions of data values with scientific significance. The graph design can be manipulated on a web browser interface, the request and manipulations steps can be stored on a personal account and the result allows installing a flux at XML or JSON format available on a web service for data mining or another indexing service.

4 OUTPUT AND SERVICES

4.1 organizing data means organizing access and improving quality

The description of data quality is an objective of the IndexMed consortium <http://www.indexmed.eu> that can be useful for data about coralligenous, based on an analysis of both similarities and differences between databases. Descriptions as metadata form a set of criteria used for data mining. The graph-based model is an abstraction tool that enables the comparison of various databases despite their differences and that improves decision support using emerging data mining methods. Practically, it is intended to give the equivalence of data, based on data dictionaries, thesauri and ontologies. From the established logical relationships, new qualifiers can be deduced including across data heterogeneity.

This work on CIGESMED data quality and their equivalence with other observatory systems involves first the analysis and description of the common elements of each piece of information, and of what differentiates them (fields name, formats, update rate, precision, observers or sensors, etc.). These descriptions are added to the data, and form a supplementary set of criteria used for data mining.
Standard formats and protocols are used to interconnect CIGESMED data with other databases. Standardization makes possible such a work, as well as a special task on interoperability qualities and accessibility of non-centralized data. It uses aggregation and new visualizations for public display, multi-interface, multi-use and multi-format, and must allow (i) the connection between many databases, and (ii) the preparation of inter-calibration works.

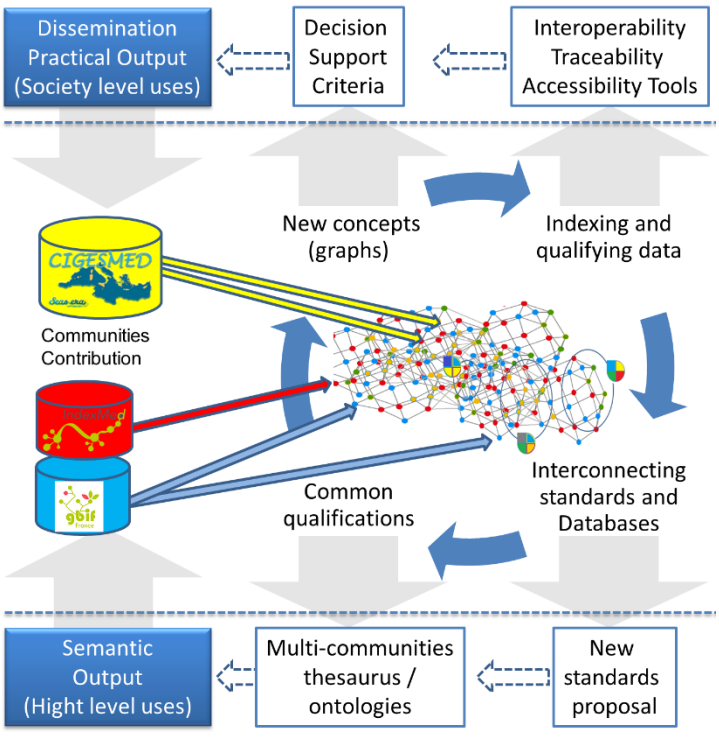


Figure 3: Iterative quality approach and IndexMed Output.

Equivalencies are used to link heterogeneous objects and construct graphs, where objects are nodes and attribute's modalities are links. The main output is a new model of dataset, stored in a graph database (graph matrix) and accessible with web-services for visualization and integrated flux. A second output is the improving of multi-community thesaurus necessary to build new common ecological concepts. The next step of this project is the recognition of patterns of context in the graph matrix that will contribute to decision criteria.

Semantic approaches greatly increase their interoperability and some initiatives using Semantic Web technologies for retrieving Biodiversity data [Amanqui et al., 2014] and developing methods for linking biodiversity and environmental

data already exist, but they often concern only an "inventory" aspect. However, it remains that specific scientific objectives, organizational logic of projects and collection of information are leading to a decentralized data distribution which may hamper environmental research development. In such a heterogeneous system footed on different organizations and data formats, not everything can be homogenized. The IndexMed workflow is a support to implement an iterative quality approach (Figure 3) increasing step by step the capacity of each data to be connected to others (i.e. contextual data like biotic, abiotic, anthropogenic and natural pressures, etc.)

The resulting cluster and their correlations to context patterns can be compared with other kinds of analyses: supervised clustering results, statistical ecology approach [Gimenez et al., 2014] and collaborative clustering methods [Forestier et al., 2008] are planned to be used at each part (e.g. a group of nearby nodes with similar patterns of context) of the graph, using job middleware (DIRAC3, [Tsaregorodtsev, 2009]). Another issue is to use "unsupervised" mode, raising the possibility to compare the results of different algorithms to achieve consensus, which acts / results in the most likely scenario. The data mining helps finding managerial values of qualifiers to propose scenarios, and provides new standardized descriptors essential for approaches such as machine learning.

4.2 Efficiency for data mining approach and links with decision support

The chain between data and decision making can be superimposed to the DIKW (u) hierarchy: Data, Information, Knowledge, Wisdom, Understanding [Zeleny, 1987; Ackoff, 1989], replacing Wisdom by Management. In a simplistic view, scientists produce knowledge by analysing data into information and by elaborating theories from information. Data constitute the primary material from which hypotheses are 1) elaborated, and 2) tested. However, even if biodiversity data have been produced in the common framework of the theory of evolution, it has often been done independently in different domains from genes to ecosystems; moreover, biodiversity data are historical in essence, they have a time component: species have been observed at a given location at a given time, and this is not reproducible as it could be done easier for physico-chemical experiments. Consequently, every piece of data in each domain may be of importance, and for older ones, they may need to be re-expressed to fit under their current conceptual and standard forms, in particular to use them all in a common approach like here. We are thus dealing with millions of pages of scientific literature and the increasing number of data repositories since Aristotle [Voultsiadou, 2007]. The Biodiversity Heritage Library <http://www.biodiversitylibrary.org/> is already making available almost 50 million pages (and still increasing), mainly up to 1930s because of copyright issues: since the scientific production progresses exponentially, we may talk here about billions of pages. Even narratives of travels and expeditions can be used to extract biodiversity semi-quantitative data [Al-Abdulrazzak et al., 2012].

The development of new data mining tools becomes crucial to explore automatically all sources of biodiversity data, or currently more reasonably semi-automatically, in order to produce the most complete knowledge that constitutes the decision support material (but not the decision-making tools by themselves!). This knowledge is the basis for developing alternative future scenarios about the biodiversity management among which decision and policy-makers will make a political choice. The graph approach may allow going a step further by integrating socio-economic knowledge in these scientifically supported scenarios. Currently, this integration is made at the decision making level, where biodiversity and socio-economics scenarios are on the contrary put in competition, most often to the benefit of the socio-economics scenarios, with too many examples from the domain of fisheries [Froese, 2011].

5 CONCLUSIONS AND RECOMMENDATIONS

Compared with dimensional and multidimensional analyses, which are often used for ecological and environmental purposes, such innovative approaches make possible the investigation of complex research questions and the emergence of new scientific hypotheses. Regarding the first results, environmental scientists and environmental managers from CIGESMED and from the IndexMed consortium have to face different challenges about links between data and well understanding of the meaning of new objects and their variation. For better analysing heterogeneously distributed data spread in different databases and for identifying statistical relationships between observed data and the emergence of contextual patterns, it will be necessary to create matches and incorporate some approximations.

The area of Decision Support Systems (DSS) focuses on development of interactive software that can use data mining export of IndexMed prototypes. This prototype aims to provide qualifiers that can be interpreted in patterns as answers to relevant decisional questions from the users, thus enhancing a person or a group to make better decisions. Till now, important efforts to develop links between Indexmed and dedicated DSS are required and possible for every particular application [Varanon et al., 2007, Power, 2007]. Specific DSS linked to IndexMed must be experimented where some successful experiences appear in several fields, like self-care management [Marschollek, 2012], water management [Pallottino, 2005], forest ecosystems [Nute, 2004] or air pollution [Oprea, 2005]. IndexMed community is open to contribution. IndexMed software are open source and privileged case studies are open data, and the involved teams plan to set up a forge and a contributory platform for expanding testing graph approaches.

Multidisciplinary approaches are a key as well as the most difficult way to improve data mining and DSS. At a “human” level, it is seriously necessary to encourage the data openness and data sharing, as the only way to give value to data after their primary use [McNutt et al., 2016]. A good start might be to organize more events dedicated to the sharing of experience and expertise, the acquisition of practical methods to construct graphs and value data through “metadata and data papers".

ACKNOWLEDGMENTS

The construction of the first prototype for IndexMed consortium is funded by the CNRS défi “VIGI-GEEK (VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel)” and CNRS INEE through the “CHARLIEE” project. Data used for this article were obtained through the European CIGESMED project <www.cigesmed.eu> (ANR conventions n° 12-SEAS-0001-01, 02 and 03 for France; GSRT - 12SEAS-12-C2 for Greece; TUBITAK Project No: 112Y393 for Turkey). The authors acknowledge the support of *France Grilles* for providing computing resources on the French National Grid Infrastructure. Supplementary acknowledgement to Gergely Sipos, Jan Bot and Roberta Piscitelli for the helpful support provided at “design your e-infrastructure” EGI <www.egi.eu> workshop. We acknowledge all the field helpers and students who have participated in data collection in the field and in the lab, as well as in data management.

REFERENCES

Ackoff, R. L., 1989. From data to wisdom. Journal of applied systems analysis, 16(1), 3-9.

Al-Abdulrazzak, D., Naidoo, R., Palomares, M. L. D., Pauly, D., 2012. Gaining perspective on what we've lost: the reliability of encoded anecdotes in historical ecology. PloS one, 7(8), e43386. doi:10.1371/journal.pone.0043386.

Amanqui, F.K., Serique, K.J., Cardoso, S.D., dos Santos, J.L., Albuquerque, A. and Moreira, D.A., 2014. Improving biodiversity data retrieval through semantic search and ontologies. Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2014 IEEE/WIC/ACM International Conference on Web Intelligence, 11-14 August 2014, Warsaw, Poland, vol.1 (WI), pp.274-281, doi: 10.1109/WI-IAT.2014.44.

Ballesteros, E., 2006. Mediterranean coralligenous assemblages: a synthesis of present knowledge. Oceanography and Marine Biology: An Annual Review, 44, 123-195.

Bianchi, C. N., Cattaneo-Vietti, R., Morri, C., Navone, A., Panzalis, P., Orru, P., 2007. Coralligenous formations in the Marine Protected Area of Tavolara Punta Coda Cavallo(N-E Sardinia, Italy). Biologia marina mediterranea, 14(2), 148-149.

Conruyt, N., Sébastien, D., Vignes-Lebbe, R., Cosadia, S., 2010. Moving from biodiversity information systems to biodiversity information services. Information and Communication Technologies for Biodiversity Conservation and Agriculture, 107-128.

Cryer, P., Hyam R., Miller C., Nicolson, N., Tuama, É.Ó, Page, R., Rees, J., Riccardi, G., Richards, K., Whitev, R., 2009. Adoption of persistent identifiers for biodiversity informatics. Report published by GBIF Secretariat.

Cortez, P., & Embrechts, M. J., 2013. Using sensitivity analysis and visualization techniques to open black box data mining models. Information Sciences, 225, 1-17.

David, R., Féral, J. P., Blanpain, C., Diaconu, C., Dias, A., Gachet,S., Gibert, K., Lecubin, J., Surace, C., 2015. A First Prototype for Indexing, Visualizing and Mining Heterogeneous Data in Mediterranean Ecology: Within the IndexMed Consortium Interdisciplinary Framework. In 2015 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS).232-239.

Deter, J., Descamp, P., Ballesta, L., Boissery, P., & Holon, F., 2012. A preliminary study toward an index based on coralligenous assemblages for the ecological status assessment of Mediterranean French coastal waters. Ecological indicators, 20, 345-352.

Deter, J., Descamp, P., Boissery, P., Ballesta, L., & Holon, F., 2012. A rapid photographic method detects depth gradient in coralligenous assemblages. Journal of Experimental Marine Biology and Ecology, 418, 75-82.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P., 1996. From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

Féral, J.-P., David, R., 2013. L'environnement, un système global dynamique. Zone côtière et développement durable, une équation à résoudre. In: Euzen, A., Eymard, L., Gaill, F. (Eds), 2013. Le développement durable à découvert, CNRS éditions: Paris, September,96-97, ISBN : 978-2-271-07896-4.

Féral, J.-P., Arvanitidis, C., Chenuil, A., Çinar, M.E., David, R., Frémaux, A., Koutsoubas, D., Sartoretto, S., 2014. CIGESMED, Coralligenous based Indicators to evaluate and monitor the « Good Environmental Status » of the MEDiterranean coastal waters, a SeasEra project (www.cigesmed.eu). Proceedings RAC/SPA 2nd Mediterranean Symposium on the Conservation of coralligenous and other calcareous bio-concretions, Portorož, Slovenia, October, 15-21.

Forestier, G., Wemmert, C., Gançarski, P., 2008. Multisource images analysis using collaborative clustering. EURASIP Journal on Advances in Signal Processing, 2008, 133.

Froese, R., 2011. Fishery reform slips through the net. Nature, 475(7354), 7-7.

Gachet, S., Véla, E., Tatoni, T., 2005. BASECO: a floristic and ecological database of Mediterranean French flora. Biodiversity & Conservation, 14(4), 1023-1034.

Gatti, G., Bianchi, C. N., Morri, C., Montefalcone, M., Sartoretto, S., 2015. Coralligenous reefs state along anthropized coasts: Application and validation of the COARSE index, based on a rapid visual assessment (RVA) approach. Ecological Indicators, 52, 567-576.

Gibert, K., Valls, A., Batet, M., 2014. Introducing semantic variables in mixed distance measures: Impact on hierarchical clustering. Knowledge and Information Systems, 40(3), 559-593.

Gibert, K., Conti, D., Vrecko, D., 2012. Assisting the end-user in the interpretation of profiles for decision support. an application to wastewater treatment plants. Environmental Engineering and Management Journal, 11(5), 931-944.

Gimenez, O., Buckland, S.T., Morgan, B.J.T., Bez, N., Bertrand, S., Choquet, R., Dray, S., Etienne, M.P., Fewster, R., Gosselin, F., Mérigot, B., Monestiez, P., Morales, J., Mortier, F., Munoz, F., Ovaskainen, O., Pavoine, S., Pradel, R., Schurr, F.M., Thomas, L., Thuiller, W., Trenkel, V., de Valpine, P., Rexstad, E., 2014. Statistical ecology comes of age. Biology letters, 10(12), 20140698.

Hong,J. S., 1982. Contribution to the study of populations of the coralligenous concretionary bottom from the Marseille region on the northwestern Mediterranean coast. Bull. Korea Ocean Res. Dev. Inst., 4: 1–2.

Kattge, J., Ogle, K., Bönisch, G., Díaz, S., Lavorel, S., Madin, J., Nadrowski, K., Nöllert, S., Sartor, K., Wirth, C., 2011. A generic structure for plant trait databases. Methods in Ecology and Evolution, 2(2), 202-213.

Kipson, S., Fourt, M., Teixidó, N., Cebrian, E., Casas, E., Ballesteros, E., Zabala, M., Garrabou, J., 2011. Rapid biodiversity assessment and monitoring method for highly diverse benthic communities: a case study of Mediterranean coralligenous outcrops. PLoS One, 6(11), e27103.

Kissling, W. D., Hardisty, A., García, E. A., Santamaria, M., De Leo, F., Pesole, G., Freyhof, J., Wissel, S., Konijn, J., Los, W., 2015. Towards global interoperability for supporting biodiversity research on essential biodiversity variables (EBVs). Biodiversity, 16(2-3), 99-107.

Klemeš, J.J. (Ed.), 2015. Assessing and measuring environmental impact and sustainability. Butterworth-Heinemann. 608 pp. ISBN: 9780127999685

Laborel, J., 1961. Le concrétionnement algal "coralligène" et son importance géomorphologique en Méditerranée. Recueil des travaux de la Station marine d'Endoume, 23(37), 37-60.

Laporte, M. A., Garnier, E., 2012. ThesauForm—Traits: A web based collaborative tool to develop a thesaurus for plant functional diversity research. Ecological Informatics, 11, 34-44.

Laporte, M. A., Mougenot, I., Garnier, E., Stahl, U., Maicher, L., Kattge, J., 2014. A semantic web faceted search system for facilitating building of biodiversity and ecosystems services. In Data Integration in the Life Sciences (pp. 50-57). Springer International Publishing.

Laubier, L., 1966. Le Coralligène des Albères. Monographie biocénotique. Adaptations chez les Annélides Polychètes interstitielles. Thès. Fac. Sci. Univ. Paris (Sér. A, No. 4693 No d'ordre 5541), 139-316.

Madin, J. S., Bowers, S., Schildhauer, M. P., Jones, M. B., 2008. Advancing ecological research with ontologies. Trends in Ecology & Evolution, 23(3), 159-168.

Marschollek, M., 2012. Decision support at home (DS@ HOME)–system architectures and requirements. BMC medical informatics and decision making, 12(1), 43.

McNutt, M., Lehnert, K., Hanson, B., Nosek, B. A., Ellison, A. M., King, J. L., 2016. Liberating field science samples and data. Science, 351(6277), 1024-1026.

Michener, W. K., Jones, M. B., 2012. Ecoinformatics: supporting ecology as a data-intensive science. Trends in ecology & evolution, 27(2), 85-93.

Noss, R. F., 1990. Indicators for monitoring biodiversity: a hierarchical approach. Conservation biology, 4(4), 355-364.

Nute, D., Potter, W. D., Maier, F., Wang, J., Twery, M., Rauscher, H. M., Knopp, P., Thomasma, S., Dass, M., Uchiyama, H., Glende, A., 2004. NED-2: an agent-based decision support system for forest ecosystem management. Environmental Modelling & Software, 19(9), 831-843.

Oprea, M., 2005. A case study of knowledge modelling in an air pollution control decision support system. Ai Communications, 18(4), 293-303.

Pallottino,S., Sechi, G.M., Zuddas, P., 2005. A DSS for water resources management under uncertainty by scenario analysis. Environmental Modelling & Software, 20 (8): 1031-1042, doi: 10.1016/j.envsoft.2004.09.012.

Pereira, H.M., Ferrier, S., Walters, M., Geller, G.N., Jongman, R.H., Scholes, R.J., Bruford, M.W., Brummitt, N., Butchart, S.H., Cardoso, A.C., Coops, N.C., Dulloo, E., Faith, D.P., Freyhof, J., Gregory, R.D., Heip, C., Höft, R., Hurtt, G., Jetz, W., Karp, D.S., McGeoch, M.A., Obura, D., Onoda, Y., Pettorelli, N., Reyers, B., Sayre, R., Scharlemann, J.P., Stuart, S.N., Turak, E., Walpole, M., Wegmann, M., 2013. Essential biodiversity variables. Science, 339 (6117): 277-278. doi: 10.1126/science.1229931

Pergent-Martini, C., Alami, S., Bonacorsi, M., Clabaut, P., Daniel, B., Ruitton, S., Sartoretto, S., Pergent, G., 2014. New data concerning the coralligenous atolls of cap corse: an attempt to shed light on their origin. RAC/SPA 2nd Mediterranean Symp. on the Conservation of coralligenous and other calcareous bio-concretions, Portorož (SIovenia), 29-30/10/2014, 129-134.

Peters, D.P., Havstad, K. M., Cushing, J., Tweedie, C., Fuentes, O., Villanueva-Rosales, N., 2014. Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology. Ecosphere, 5(6), 1-15.

Power,D.J., 2007. A Brief History of Decision Support Systems, DSSResources.COM (Editor), World Wide Web, version 4.0", March 2007, http://dssresources.com/history/ dsshistory.html.

Reichman, O. J., Jones, M. B., Schildhauer, M. P., 2011. Challenges and opportunities of open data in ecology. Science, 331(6018).

Sartoretto, S., 1994. Structure et dynamique d'un nouveau type de bioconstruction à Mesophyllum lichenoides (Ellis) Lemoine (Corallinales, Rhodophyta). Comptes rendus de l'Académie des sciences. Série 3, Sciences de la vie, 317(2), 156-160.

Sartoretto,S., Schohn, T., Bianchi, C.N., Morri, M.C., Garrabou, J., Ballesteros, E., Ruitton, S., Verlaque, M., Daniel, B., Charbonnel, E., Blouet, S., David, R., Féral, J.-P., Gatti, G., 2016. An integrated approach to evaluate and monitor the conservation state of coralligenous habitats: the Index-Cor approach. submitted in Ecological Indicators.

Sini, M., Kipson, S., Linares, C., Koutsoubas, D., Garrabou, J., 2015. The Yellow Gorgonian Eunicella cavolini: demography and disturbance levels across the Mediterranean Sea. PloS one, 10(5), e0126253.

Tsaregorodtsev, A., 2009. DIRAC3 . The New Generation of the LHCb Grid Software. Journal of Physics: Conference Series, vol. 219 062029, n° 6.

Trygonis, V., Sini, M., 2012. photoQuad: A dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. Journal of Experimental Marine Biology and Ecology, 424, 99-108.

Uraikul, V., Chan, C. W., Tontiwachwuthikul, P., 2007. Artificial intelligence for monitoring and supervisory control of process systems. Engineering Applications of Artificial Intelligence, 20(2), 115-131.

Voultsiadou, E., 2007. Sponges: An historical survey of their knowledge in Greek antiquity. Journal of the Marine Biological Association of the United Kingdom, 87(06), 1757-1763.

Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T., Vieglais, D., 2012. Darwin Core: An evolving community-developed biodiversity data standard. PloS one, 7(1), e29715.

Zeleny, M., 1987. Management support systems: towards integrated knowledge management. Human systems management, 7 (1): 59-70.