SITIS 2015, 11th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), 23-27 nov. 2015, Bangkok, Thailand, pp. 232-239, DOI:10.1109/SITIS.2015.119

A first prototype for indexing, visualizing and mining heterogeneous data in Mediterranean ecology

within the IndexMed consortium interdisciplinary framework

R. DAVID, J.-P. FERAL, S. GACHET, A. DIAS
Institut Méditerranéen de Biodiversité et d'Ecologie marine et continentale (IMBE)
CNRS, Aix Marseille Université
IRD, Université d'Avignon
Marseille, FRANCE
romain.david@imbe.fr, jean-pierre.feral@imbe.fr, sophie.gachet@imbe.fr, alrick.dias@imbe.fr

C. BLANPAIN, J. LECUBIN
Service informatique (SIP)
OSU Pythéas, CNRS
Aix Marseille Université
Marseille, FRANCE
cyrille.blanpain@osupytheas.fr, julien.lecubin@osupytheas.fr

C. DIACONU
Centre de Physique des Particules, CNRS
Aix Marseille Université
Marseille, FRANCE
diaconu@cppm.in2p3.fr

C. SURACE
Laboratoire d'Astrophysique de Marseille (LAM)
CNRS, Aix Marseille Université
Marseille, FRANCE
christian.surace@lam.fr

K. GIBERT
Department of Statistics and Operations Research
Universitat Politecnica de Catalunya
Barcelona, SPAIN
karina.gibert@upc.edu

***Abstract*—: Although biodiversity has been extensively studied over the last centuries, recent evidences suggest that links between collected data are still be missing. In order to fill this knowledge gap and at the initiative of the CNRS Institute of Ecology and Environment (INEE), IndexMed <www.indexmed.eu>, a unique and multidisciplinary consortium consisting of ecologists, sociologists, economists, mathematicians, IT specialists and astronomers was created. Through exploratory projects, IndexMed develops new methods for analyzing data on Mediterranean biodiversity and implements solutions based on interoperability technologies already deployed in other disciplines. In particular, IndexMed aims to build a prototype of such data graphs and "data cubes". This paper will first explore the ability of tools and methods by means of graphs to connect biodiversity objects with non-centralized data. It will then introduce the use of algorithms and graphs to analyze environmental and societal responses and presents a prototype under development.**

Indexing data, data qualification, data traceability, decentralized information systems, data-mining, ecology

I. INTRODUCTION

Although considered by most scientific disciplines and industries producing and using information as the most promising opportunity for progress and discoveries, the use of Big Data in Ecology is still lagging behind [34]. Information systems linking objects with qualifications (links) are omnipresent, the first object being the consumer. All these links can be used for data mining process. Data mining is the computational process allowing discovering patterns in large data sets («big data») and involves methods at the crossroads of artificial intelligence, machine learning, statistics, and databasing systems. (<www.kdd.org/>). Nowadays, many businesses have understood its huge analytical potential (insurance to manage risk, games companies to find cheaters, banks for investments, traders to increase their margins, advertisers and networks to increase their social/commercial impacts, etc.) Today, the environmental emergency requires a response through a shared system connected to local and global issues and beyond scientific questions such as: "Is this degradation related to this particular pressure?" instead addressing the following question: "How can we improve/preserve the ecological condition of an environment in the most efficient way?". Such queries help identify the limits not to be exceeded, for a set of conditions which may have opposite or complementary effects.

Data in ecology are extremely heterogeneous (Fig. 1). To get access to the different type of data and data formats, the system must be distributed, *i.e.* data remain where they were produced, although normalized flows are deployed and can be accessed by all members of the network, both at local and international levels and with an index to list and describe each record based on shared typologies. There are already some initiatives using Semantic Web technologies for retrieving Biodiversity data [1], Improve knowledge and developing methods for linking of biodiversity and environmental data, but they often concern only an "inventory" aspect of biodiversity (collection, observations, repositories and distribution) and far less functional aspects (Catalog of life, Data-ONE [Data Observation Network for Earth], EMODnet [European Marine Observation and Data Network], GEO and EU-BON [Group On Earth Observations Biodiversity Observation Network] and [European Biodiversity

Observation Network], GBIF [Global Biodiversity Information Facilities], LifeWatch, OBIS [Ocean Biogeographic Information System], TDWG [Biodiversity Information Standards] - with Darwin Core and ABCD - are well-known examples of efforts for interoperability and standardizing data collection regarding biodiversity).

The newly created IndexMed consortium has set a goal to build such a distributed system, with the help of institutes from different disciplines, in order to overcome skills lacking among ecologists.

IndexMed was created by the axis "Management of biodiversity and natural spaces" of the IMBE (Mediterranean Institute of Biodiversity and marine and terrestrial Ecology) with the aim to develop the knowledge of databases and their effective use in the ecological research community.

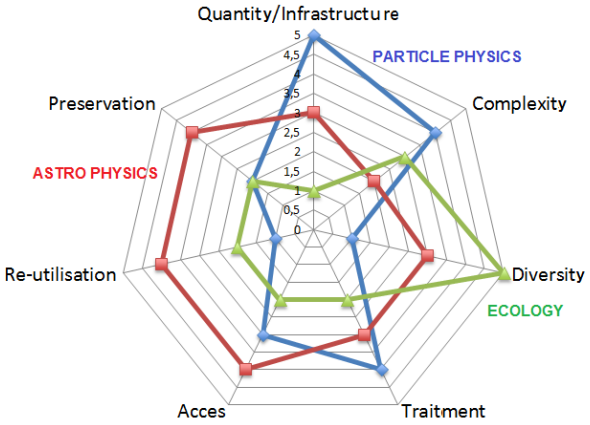

Figure 1. First summary of drawn conclusions in the MASTODONS (very large scientific datasets) meeting. Big Data in ecology are compared to astrophysics and particle physics data and demonstrate that diversity is the most important aspect to consider when dealing with such data.

In particular, this consortium responds to project calls and uses databases and address ecological issues in the Mediterranean Basin, therefore promoting multidisciplinary and collaboration across CNRS [National Centre for Scientific Research] institutes, other research entities and universities. The projects developed by members of the IndexMed consortium must be based on various national and international initiatives and promote international collaborations, therefore connecting existing networks to initiatives at national and international levels.

IndexMed's short-term goal is to establish a platform indexing Mediterranean biodiversity data and environmental parameters which are of interest for many researchers. This index will employ the tools and methods recommended 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, Life-Watch, GEO-BON, etc.) along with other research entities (i.e. IRD [Institute of Research for the Development] or MNHN [National Museum of Natural History]).

The architecture of the proposed prototype of information systems for projects being developed is decentralized, and is a first step towards indexing, classifying, mapping and interfacing data from coastal and marine Mediterranean environments, essential in ecology research and natural spaces management tools.

Building on the efficiency of this prototype, the project will develop an "object resolution service" (*i.e.* a web service that finds links and dependencies among indexed objects, based on unique objects identification). This object resolution service should allow an inventory of biodiversity descriptors, estimate the capacity of data to describe socio-ecological systems at temporal and geographic scales, and must permit links between economic and social approaches. It is based on large panels of participants, data and skills being developed. The project should also develop new trans-disciplinary methods of data analysis, focusing on open data, open source and free methods and development tools.

II. METHOD

A. Decentralized information systems

Integrating databases into one large centralized database has been attempted by many programs, but always failed. Then, we must ensure our ecological, economics and social sciences data are interoperable, connectable and comparable, and to manage to analyze them without placing one above the other.

The principle of information systems (IS) decentralization is unavoidable once one looks at the real-time data analysis produced by different actors in various fields. Whether it is used for biodiversity studies or for the knowledge of socio-ecological systems, the production of data is expensive and rarely automated. The long time series and/or large spatial extent studies are difficult to conduct, and when dealing to "interpretive data" the use of too many observers affect the reliability of the observation.

Reproducibility today is frequently questioned or even refuted. In the context of multi-source data production, it is essential to help each producer as well as external users to install and maintain a suited IS for their needs (maintenance, development of software, developments on the data scheme, scope and standards of data). This decentralization requires working on data models, their evolution, but must remain consistent with the data collection protocols and their evolution.

A "modular" organization (for administering a type of object or data independently by the most competent actor) is preferred to centralized systems: in this interdisciplinary framework, based on systems observing at large scale, each participant cannot consolidate data from all disciplines. The data serve as a model and concern marine habitat and common terrestrial habitat for all disciplines. This type of methodology may be declined on many environmental models, including but not limited to terrestrial and marine habitats, organisms, communities and species assemblages.

Standardization makes possible such a work, as well as a special task on interoperability qualities and accessibility of non-centralized data. It use aggregation for public display, multi-interface, multi-use and multi-format, and must permit (i) the connection between many databases, and (ii) the preparation of inter-calibration works.

B. Indexing data

Indexing is an alternative to centralization, which allows global approaches in ecology. Global ecology considers ecological systems and their complexity in terms of composition, structure and interactions. It identifies the parameters able to lay the foundations for sustainable management of resources and services they provide, to better understand and anticipate the risks and their consequences, and to participate in the improvement of the quality of life of societies.

The IS data is the keystone of the linkage of different databases formats. It provides unique identifiers for data objects which can return metadata about themselves and which can be brought together into a distributed collaborative information system as recommended by GBIF [6]. It identifies each record, recordings of each state (version), and solving for each of these states all the data that have been used and their condition, in order to reach a new transformed state. It allows describing all the previous states, thus ensuring provenance, traceability and intellectual property of this data if it exists, to identify the adjectives that this new data can or cannot inherit, and thus it complements criteria that may serve as an additional descriptor for data mining. The resolution of these indexed web service relies on the unique identifiers of databases, creates some where none already exist, and also creates relationships between them where there is more than one.

This Information System allows making an informed choice of data aggregation as nodes, because it does not contain data considered as sensitive by a data producer. These indexing nodes will be "clonable" on a discretionary basis with enrichment rules and sharing licenses corresponding to "creative common" type "sharing the same conditions", allowing others to copy, distribute and modify the index (Fig 2), provided they publish express any adaptation of the index under the same conditions (open-source, open data). These rules will encourage the emergence of standards to improve the interoperability of data and promote the participation of new contributor laboratories taking into account their contribution to the technical possibilities as and when the project develops.

The prerequisites of these tasks are accessibility of data, normalized and qualified flow with open data accessible by means of fluxes. Tools to be improved are resolution service, unique identifiers, equivalence between identifiers, and reproducible indexing nodes. The main objective is availability and traceability of data; it should permit a heritability for data qualifiers in the future, stories of data / data life studies in ecology and associated disciplinary, mainly economics, sociology and law.

C. Quality and qualification of data

In a production framework of multi-source data, the equivalence of observation system's problematics and inter-calibration of observers then become crucial. Increasingly, the need for integrative multi- or trans-disciplinary approaches becomes necessary, in the study of systems where data output in each discipline is discontinuous, imprecise and badly distributed.

Yet all the variables of these systems interact in time and at each spatial scale (biotic, abiotic variables, anthropogenic and natural pressures, perceived and provided services, societal perception, etc.) [4] [9] [13]. To prepare a true integrated management biodiversity [10], the data collection protocols must allow a fine and standardized description with a shared typology about (i) conditions of measurement and ecological status, (ii) pressures on these areas, and (iii) economic and sociological context. All this knowledge is necessary and should be standardized on a large scale to enable decision support.

Scientific challenges about data quality are complicated, given (i) their volume and the dynamics of their update, the update repositories and the standards necessary for administering the data, (ii) their intrinsic heterogeneity and complexity, especially related to cross biodiversity data and contextual variables, (iii) the heterogeneity of users, networks of producers actors and their motivations to maintain and supply their information systems.

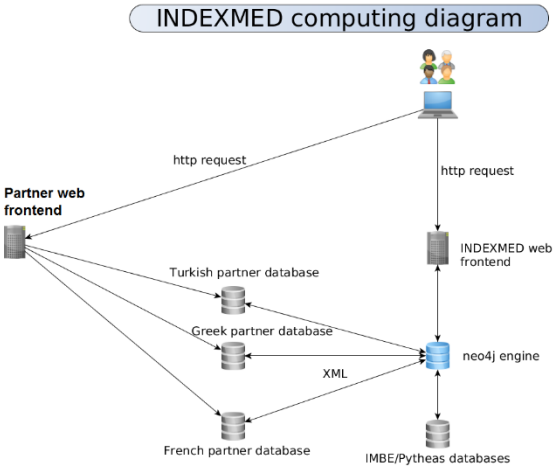

Figure 2. Global hardware architecture: it provides graph results from local or remote databases. Each partner will store data locally and the system will be able to get results using standard protocols. Graph results can be obtained by end users using a specific web frontend. 3 countries are testing the architecture under European programs and inventories of marine biodiversity (CIGESMED, DEVOTES, ZNIEFF)
In addition, queries on remote databases are being tested (*e.g.* GBIF (species occurrences), Tela Botanica (French Flora), European Pollen databases. Other thematic databases will be integrated in the development of this project.

Work on data quality and their equivalence is required. It firstly involves the analysis and description of the common elements of each piece of information, and what differentiates them (name fields, formats, update rate, precision, observers or sensors, etc.) These descriptions are added to the data and form a body of criteria used for data mining. Secondly, it is intended to give the equivalence of data, based on data dictionaries and thesaurus. Some database conjunctions allow to deduce other, using firstly their own ontology for each domain and multidisciplinary. Out of all these logical relationships, we can deduce new qualifiers that are either new data quality or a way to find common qualifier to heterogeneous data that can serve as an additional descriptor as part of the data mining.

This work is possible by accessible and decentralized IS, commonly related by other systems and normalized, accessible and configurable stream of qualifications. These standards/data dictionary/thesaurus ontology are

commonly constructed and will be improved over the next years in an open database. This work will allow jointed analysis of different data corpus, and inter-calibrations of data productions.

III. First Results

A. Data visualization and graphs

The visualization prototype is initiated in the VIGI-GEEK project (VIsualisation of Graph In transdisciplinary Global Ecology, Economy and Sociology data-Kernel). It currently involves representations of graphs, but in the coming years it will explore other types of heterogeneous data representations.

A major goal is to produce a "multidisciplinary" tool for construction and graph visualization through IndexMed. These graphs are constructed from aggregated information through the nodes indexing and qualifying data concerning Mediterranean marine and coastal environment, in various disciplines (socio-ecology, econometrics, ecology [structural and operational], town planning, management, etc.) at the Mediterranean scale. The development of this prototype is to make customizable graphs to search and visualize multidisciplinary data putting on the same level socio-ecological, economical, ecological, molecular and functional data types (trophic relationships, functional traits, etc.) Each object or concept describing biodiversity can be a node, and the terms of the attributes that describe them are as many links which can give specific properties (attraction or repulsion of other objects). Particular graphs can mix heterogeneous objects, if these different objects share identical terms for any attributes (Fig. 3).

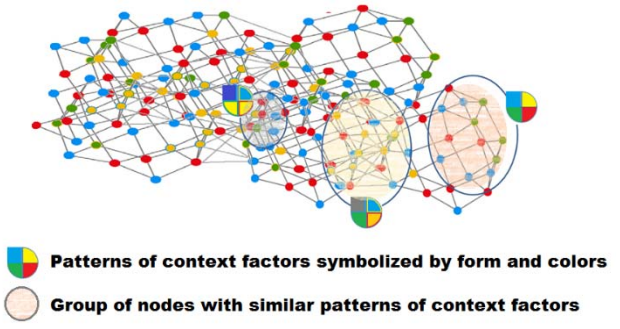

Figure 3. IndexMed project graphic sample using IMBE dataset: different objects (3 in this case: species, quadrat photo samplings and localities) can be linked if they share the same attribute modality. This interface allows a generic integration of different objects when links are possible with common attribute values. This enables the design of graphs which are then analyzed thanks to the stream generated in JSON or XML. This flow will be operated by the algorithms selected by a decision tree, depending on the type of objects generated graph (not yet developed). The computing infrastructure will be used to browse these graphs.

The interface use Neo4j <neo4j.com/>, a graph database implemented in java and released in 2010. The community edition of the database is licensed under the free GNU General Public License (GPL) v3. The database and its additional modules (online backup or high availability) are available under a commercial license.

Neo4j is used to represent data as objects connected by a set of relations, each object having its own properties. When the database is requested, a graph appears and it is possible to interact with it, using the web browser.

In Neo4j, everything is stored in form of either edge, node or attribute. Each node and edge can have any number of attributes. Both the nodes and the edges can be labelled and colored.

IndexMed technical staff is developing a specific web frontend using Ajax/Jquery language. It may be possible to request a database asking for specific objects and specific relations between them, without using a technical query language such as SQL or Cypher. The prototype is developed for a generic and is enough to integrate any type of data in the form of "object, attribute, and attribute value" (Fig. 4).

This prototype will be available as an open-source format to develop, on the medium term, usage of these graphs for decision support in environmental management and as part of a research project to be submitted to European calls for projects (BiodivERsA, ERDF, SeasEra, H2020…)

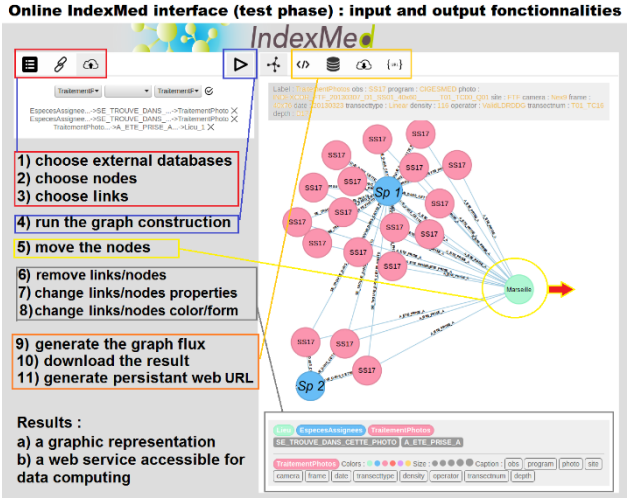

Figure 4. Data visualization coming from our specific web frontend: different types of objects can be mixed, and relation between them is due to the value of terms which describe them. Links and nodes can be selected or colored by values of term. Fluxes content and generated request are display in the front of this interface. At this state, databases on marine benthic environment, terrestrial flora and pollen databases are tested. Others data content (dendrology, archeology, economic and social data) will be included after works on common thesaurus and attributes value (which permit to construct links).

B. Testing other data devices: the astronomy example ("CHARLIEE" project with IndexMed consortium)

Astronomy is a good example of interoperability of data and associated service. Since the first works of Messier and the NGC catalogues (first typology of astronomic objects in 1888), all the extragalactic astronomical objects have been identified and have been associated with a unique identifier. Each new object is defined with a specific ID, and with a combination of its position in the sky at a given time and epoch of the observation.

In 2002 the International Virtual Observatory Alliance (IVOA <www.ivoa.net>) has been created to gather efforts on data standardization and dissemination. Since then, the Virtual Observatory (VO) allowed to spread validated data

all over the world and to use data from everywhere from earth. Infrastructure, standards and tools have been developed to easily search for data and use and export all structured objects. From Solar system and stellar objects to galactic objects, theoretical data and tabular data are covering several quantities such as astrometric, photometric, spectroscopic data. Most of the characteristics of the objects have been characterized and formatted. Format exchanges have been described as data models and serialized as XML formatted data. By the same time, access protocols have been setup to access images, spectra, tabular and temporal data. For several years, software is being developed to facilitate the cross-use and discovery of astronomical data. Among the most used software, one can cite:
ALADIN (http://aladin.u-strasbg.fr/)
TOPCAT (http://www.star.bris.ac.uk/~mbt/topcat/)
VOSPEC (http://www.sciops.esa.int/index.php?project=SAT&page=vospec)
Such tools hide the complexity of the VO infrastructure, to facilitate the use of data.

The goal is to transcript the ecological data into VO formatted data in order to use the existing astronomical tools in the study. Data will be translated using Unified Content Descriptors (UCDs). This will allow the use of astronomical tools in a comparative way. For example we can use the density maps (Fig 5).

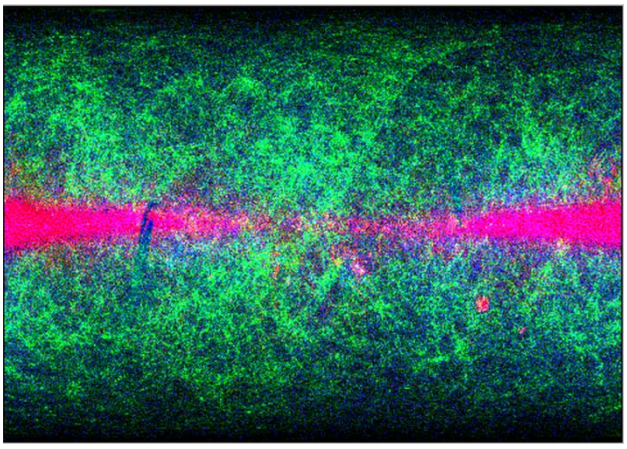

Figure 5. TOPCAT density map. The intensity of color at each pixel reflects the number of points that fall within its bounds. TOPCAT is an interactive graphical viewer and editor for tabular data. It presently provides most of the facilities that astronomers need for analysis and manipulation of source catalogues and other tables. This is very useful for analysis and visualization of extremely crowded plots. In the IndexMed framework, TOPCAT will be used for ecological, economical and/or sociological data visual representation. (http://www.starlink.ac.uk/topcat/)

IV. DISCUSSION AND PERSPECTIVES

In such an *environmental* framework, the development of a network that integrates researchers in Humanities and Social Sciences across the Mediterranean basin is essential. Indeed, the question of the compatibility of models and data from different disciplines is a key issue for building models and measuring relevant data within each discipline which follows logics and different rules. Some workshops will build-up protocols compatibility between models and ecological, social and economic data, as well as to identify priority actions for data access to be inserted in the graphs. Major tasks to be carried out after indexing and building the first graphs are to (i) build ontology and their implementations in existing graphs, (ii) create data mining tools with heterogeneous data in ecology and (iii) build a bridge between artificial intelligence tools used to analyze graphs and decision support. The decisions support objectives of the project also need to consider ways of preserving the short and long term data and thus the generated qualifiers.

A. Ontologies for complex graphs approaches

Ontology is a specification of a conceptualization for a domain of knowledge [18] and therefore it results of a choice to formally describe this area, based on a controlled vocabulary, creating dependencies and inheritance between these concepts.

Biodiversity is a multiple field of knowledge where concepts and data abound [28]. For example, it has recently been described in GEO-BON as 22 essential variables [33]. Translation into ontologies is essential for a better qualification of these data [25], to show new links between different objects that make it up (housing, key-stone species, predator species, etc.) These links give additional properties and dimensions to graphs constructed under IndexMed and increase the perimeters of the investigation areas. The use of ontologies within IndexMed project will build on existing eco-informatics and ongoing work related to thesaurus [20] [22] [42].

B. Data mining tools with heterogeneous data in ecology

Data mining emerged in the late 90s [8] as a discipline to extract relevant, novel and understandable knowledge from the analysis of datasets, easier to record, but that refer to increasingly complex phenomenon, among which we find ecology.

A main issue in this field is the need to analyze the relationships between heterogeneous data for a proper understanding of reality. The data mining on heterogeneous data is complex and convoluted by a longer work on quality data. However, it must allow the extraction of relevant decisional knowledge or acquaintance from large amounts of data, using supervised or fully automatic algorithms.

Acquiring a better global understanding of the balance of Socio-Ecological Systems (SES) and their impact on biodiversity is one of the main current scientific challenge. Advances in this understanding process must consider the construction and testing of methods for joint interpretation of these heterogeneous data.

Some research systems started to present logical inter-dependencies in SES for facilitating building of biodiversity and ecosystems services [23]. New opportunities are created by open data formats in ecology [38] and qualification standards usable in data mining are developed with the Taxonomic Databases Working Group consortium <www.tdwg.org> by the Darwin Core Task Group [43]. Some works focus on the integration of declarative knowledge with numerical and qualitative data [14] or on the post-process of results required to provide understandable knowledge to the end-user [5] [15]. Data mining methods must be able to bring new perspectives to the disciplinary research on these complex systems, studying ultimately interrelated objects (environmental chemistry, genomics, transcriptomic, proteomics, metabolomics, stands ecology, socio-ecological systems, or landscape ecology are some examples) as well as

dealing with the intrinsic space-time of the ecological phenomenon. It will use indexed data, data qualification, and data traceability for discovering patterns in the data values conjunctions with scientific significance. In the IndexMed Project, supervised clustering, graph algorithms, statistical ecology [17] and collaborative clustering methods [12] are planned to be used. 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 such as scenarios, and provides standardized descriptors essential for approaches such as machine learning. The ambition of IndexMed consortium is to achieve operational objectives in terms of decision support, based on the exploitation of these complex multidisciplinary graphs. Scientific challenges concerning the quality of data are complicated by the heterogeneity of sometimes contradictory norms and standards between disciplines. However, adopting too many standards can prevent the pooling of heterogeneous objects. A process of simplification and approximation will probably be necessary.

The ongoing more advanced initiatives concerning open linked data construct analyses with data mostly linked by geographical approaches (see [40]). One challenge raised by the consortium IndexMed consists to achieve common ontologies between several disciplines, and to offer new dimensions of analysis beyond the usual geographical and time fields (see [1]), usable by intelligent decision support machine.

C. From artificial intelligence to intelligent decision support

The area of Decision Support Systems (DSS) focuses on development of interactive software that can analyze data and provides answers to relevant decisional questions from the users, thus enhancing a person or group to make better decisions. Early DSS [24] used simple monitoring; later, model-based simulation introduced what-if analyses. Intelligent DSS (IDSS) [26] included specific domain knowledge and automatic reasoning capabilities. Till now, important efforts to develop dedicated (I)DSS are required for every particular application [41][36] and some successful experiences appear in several fields, like self-care management [27], water management [32], forest ecosystems [30] or air pollution [31]. However, upgrading of these platforms to incorporate new risk factors control, new sensors connections or to take into account new predictive models becomes costly and time consuming.

New generation IDSS provides sufficient integration for achieving a really holistic approach (taking into account not only monitoring of isolated parameters, but also information from the different data sources available, activities developed in the community and all types of available information (images, measures, qualitative data, knowledge, documents, tweet, etc.) and to get a sufficiently flexible DSS system architecture to make easy adaptation of the system to advances in the ecological state of the art, and new architectures must be designed to permit flexible upgrades or domain-changes of these kind of platforms in an easy way [3] [29]. Today, it seems clear [35] [37] [16] that IDSS must combine data-driven, analytical and knowledge-based models (including prior declarative expert knowledge), as well as some standardized reasoning [21] [19] [39] to provide a relevant support to managers, even if there are not yet much real experiences on-going under this approach.

Ecology belongs to a set of critical domains where wrong decisions may have tragic consequences. Decision-making performed by IEDSSs should be collaborative, not adversarial, not only finding optimal or suboptimal solutions, but making the entire process more open and transparent. The system will have to deal with inherent uncertainty of decisions and decisions must inform and involve those who must live with the (consequences of the) decisions.

D. Preservation of heterogeneous data

The scientific data is collected with large material and human efforts and tend to be unique due to the ever increasing experimental complexity and because of the time-stamped nature of the data itself, as it is very often in ecology studies. The data preservation is therefore of crucial importance and may open new avenues for research at low cost, and for instance when older data sets are combined with newer data in order to enhance the precision or to detect time-dependent variations. Various disciplines have organized the preservation of larger or smaller samples of experiments in a different manner, most of the time adjusted to the community needs [6]. However, there is a large potential for improvement and unless an immediate and massive action is started, the danger to simply loosing unique data sets remains significant. Investigations in a multi-disciplinary context have revealed many similarities across heterogeneous data sets within some scientific disciplines; for instance, several experiments in high-energy physics may choose to standardize their preservation approaches [2] and astrophysics data is massively mutualized in the IVOA. Moreover, there are similarities and constructive complementarities in the scientific needs and methods for data preservation across different scientific disciplines. These arguments suggest that a rigorous approach of the heterogeneous data treatment, from collection, treatment and access to mining, visualization and storage can also be beneficial for the long term preservation. In addition, the preservation of data sets already collected can be reorganized through novel algorithms in order to enhance the robustness of the data preservation systems, thereby extracting more science and enhancing the original investment in experiments and data collection. Due to these considerations, this is one of IndexMed Consortium goals.

ACKNOWLEDGMENT

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 with the "CHARLIEE" project. Data used for this article was obtained through the CIGESMED project [11] <www.cigesmed.eu>.

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. Thanks are also due to Dr D. Vauzour for improving the English text.

REFERENCES

[1] F.K. Amanqui, K.J. Serique, S.D. Cardoso, J.L. dos Santos, A. Albuquerque and D.A. Moreira, "Improving Biodiversity Data Retrieval through Semantic Search and Ontologies," in 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

[2] Z. Akopov, S. Amerio, D. Asner, E. Avetisyan, O. Barring, J. Beacham, M. Bellis, G. Bernardi, S. Bethke, A. Boehnlein, T. Brooks, T. Browder, R. Brun, C. Cartaro, M. Cattaneo, G. Chen, D. Corney, K. Cranmer, R. Culbertson, S. Dallmeier-Tiessen, D. Denisov, C. Diaconu, V. Dodonov, T. Doyle, G. Dubois-Felsmann, M. Ernst, M. Gasthuber, A. Geiser, F. Gianotti, P. Giubellino, A. Golutvin, J. Gordon, V. Guelzow, T. Hara, H. Hayashii, A. Heiss, F. Hemmer, F. Hernandez, G. Heyes, A. Holzner, P. Igo-Kemenes, T. Iijima, J. Incandela, R. Jones, Y. Kemp, K. Kleese van Dam, J. Knobloch, D. Kreincik, K. Lassila-Perini, F. Le Diberder, S. Levonian, A. Levy, Q. Li, B. Lobodzinski, M. Maggi, J. Malka, S. Mele, R. Mount, H. Neal, J. Olsson, D. Ozerov, L. Piilonen, G. Punzi, K. Regimbal, D. Riley, M. Roney, R. Roser, T. Ruf, Y. Sakai, T. Sasaki, G. Schnell, M. Schroeder, Y. Schutz, J. Shiers, T. Smith, R. Snider, D.M. South, R. St. Denis, M. Steder, J. Van Wezel, E. Varnes, M. Votava, Y. Wang, D. Weygand, V. White, K. Wichmann, S. Wolbers, M. Yamauchi, I. Yavin, H. von der Schmitt [DPHEP Study Group]. "Status Report of the: Towards a Global Effort for Sustainable Data Preservation in High Energy Physics", DPHEP Study Group Collaboration, May 2012, 93 pp., DPHEP-2012-001, FERMILAB-PUB-12-878-PPD, e-Print: arXiv: 1205.4667 [hep-ex].

[3] O.J. Bott, M. Marschollek, K.H. Wolf and R. Haux, "Towards new scopes: sensorenhanced regional health information systems - part 1: architectural challenges". Methods of Informations in Medecine vol. 46(4), 2007, pp. 476–483.

[4] N. Conruyt, D. Sébastien, S. Cosadia, R. Vignes-Lebbe and T. Touraivane, "Moving from biodiversity information systems to biodiversity information services". In: Information and Communication Technologies for Biodiversity, Conservation and Agriculture, L. Maurer and K. Tochtermann (Eds.). Shaker Verlag: Aachen, August 2010.

[5] P. Cortez and M.J. Embrechts "Using sensitivity analysis and visualization techniques to open black box data mining models". Information Sciences, vol. 225, March 2013, pp. 1-17, doi:10.1016/j.ins.2012.10.039

[6] P. Cryer, R. Hyam, C. Miller, N. Nicolson, É.Ó Tuama, R. Page, J. Rees, G. Riccardi, K. Richards, and R. White. "Adoption of persistent identifiers for biodiversity informatics: Recommendations of the GBIF LSID GUID task group", 6 November 2009. Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark (version 1.1, last updated 21 Jan 2010) 62.

[7] C. Diaconu [Ed.]. PREDON group, "Scientific Data Preservation 2014", February 2014, 61pp., http://predon.org

[8] U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, "From Data Mining to Knowledge Discovery: An overview". In Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, Fall 1996, pp.37-54.

[9] J.-P. Féral [Ed.], "Concepts and methods for studying marine biodiversity, from gene to ecosystem". Océanis, documents océanographiques, vol. 24(4) [1998], Institut Océanographique, Paris, March 2001, 420 pp., ISSN: 0182-0745.

[10] J.-P. Féral and R. David, "L'environnement, un système global dynamique. - 22. Zone côtière et développement durable, une équation à résoudre". In: Le développement durable à découvert, A. Euzen, L. Eymard, F. Gaill (Eds), CNRS éditions: Paris, September 2013, pp. 96-97, ISBN : 978-2-271-07896-4

[11] J.-P. Féral, C. Arvanitidis, A. Chenuil, M.E. Çinar, R. David, A. Frémaux, D. Koutsoubas and S. Sartoretto, "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 2014, pp. 15-21

[12] G. Forestier, C. Wemmert and P. Gançarski, "Multi-source Images Analysis Using Collaborative Clustering". EURASIP Journal on Advances in Signal Processing, Special issue on Machine Learning in Image Processing, 2008, Article ID 374095, 11 pp., doi:10.1155/2008/374095

[13] S. Gachet, E. Véla and T. Tatoni, "BASECO: a floristic and ecological database of Mediterranean French flora". Biodiversity and Conservation, vol. 14, April 2005, pp.1023-1034, doi: 10.1007/s10531-004-8411-5

[14] K. Gibert, A. Valls and M. Batet, "Introducing semantic variables in mixed distance measures. Impact on hierarchical clustering", Knowledge and Information Systems, vol. 40(3), September 2014, pp. 559-593, doi: 10.1007/s10115-013-0663-5

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

[16] K. Gibert, M. Sánchez-Marrè and I. Rodríguez-Roda, "GESCON-DA: An intelligent data analysis system for knowledge discovery and management in environemntal databases". Environmental modelling and software, vol. 21(1), January 2006, pp. 115-120, doi:10.1016/j.envsoft. 2005.01.004

[17] O. Gimenez, S.T. Buckland, B.J.T. Morgan, N. Bez, S. Bertrand, R. Choquet, S. Dray, M.P. Etienne, R. Fewster, F. Gosselin, B. Mérigot, P. Monestiez, J. Morales, F. Mortier, F. Munoz, O. Ovaskainen, S. Pavoine, R. Pradel, F.M. Schurr, L. Thomas, W., Thuiller, V. Trenkel, P. de Valpine and E. Rexstad, "Statistical ecology comes of age". Biology Letters, vol. 10, December 2014, 4 pp., doi: 10.1098/ rsbl.2014.0698.

[18] T.R. Gruber, "A translation approach to portable ontology specifications". Knowledge Acquisition, vol 5(2), June 1993, pp. 199-220, http://dx.doi.org/10.1006/knac.1993.1008.

[19] A. Helmer, B. Song, W. Ludwig, M. Schulze, M. Eichelberg, A. Hein, U. Tegtbur, R. Kayser, R. Haux and M. Marschollek, "A sensor-enhanced health information system to support automatically controlled exercise training of COPD patients". In: 4th International Conference on Pervasive Computing Technologies for Healthcare. Munich: IEEE, 22-25 March 2010, pp. 1-6, doi: 10.4108/ CSTPERVASIVEHEALTH2010.8827

[20] J. Kattge, S. Díaz, S. Lavorel, I.C. Prentice, P. Leadley, G. Bönisch, E. Garnier, M. Westoby, P.B. Reich, I.J. Wright, J.H.C. Cornelissen, C. Violle, S.P. Harrison, P.M. van Bodegom, M. Reichstein, N.A. Soudzilovskaia, D.D. Ackerly, M. Anand, O. Atkin, M. Bahn, T.R. Baker, D. Baldocchi, R. Bekker, C. Blanco, B. Blonder, W. Bond, R. Bradstock, D.E. Bunker, F. Casanoves, J. Cavender-Bares, J. Chambers, F.S.I. Chapin, J. Chave, D. Coomes, W.K. Cornwell, J.M. Craine, B.H. Dobrin, W. Durka, J. Elser, B.J. Enquist, G. Esser, M. Estiarte, W.F. Fagan, J. Fang, F. Fernández, A. Fidelis, B. Finegan, O. Flores, H. Ford, D. Frank, G.T. Freschet, N.M. Fyllas, R. Gallagher, W. Green, A.G. Gutierrez, T. Hickler, S. Higgins, J.G. Hodgson, A. Jalili, S. Jansen, A.J. Kerkhoff, D. Kirkup, K. Kitajima, M. Kleyer, S. Klotz, J.M.H. Knops, K. Kramer, I. Kühn, H. Kurokawa, D. Laughlin, T.D. Lee, M. Leishman, F. Lens, T. Lenz, S.L. Lewis, J. Lloyd, J. Llusià, F. Louault, S. Ma, M.D. Mahecha, P. Manning, T. Massad, B. Medlyn, J. Messier, A. Moles, S. Müller, K. Nadrowski, S. Naeem, Ü. Niinemets, S. Nöllert, A. Nüske, R. Ogaya, J. Oleksyn, V.G. Onipchenko, Y. Onoda, J. Ordoñez, G. Overbeck, W. Ozinga, S. Patiño, S. Paula, J.G. Pausas, J. Peñuelas, O.L.

Phillips, V. Pillar, H. Poorter, L. Poorter, P. Poschlod, R. Proulx, A. Rammig, S. Reinsch, B. Reu, L. Sack, B. Salgado, J. Sardans, S. Shiodera, B. Shipley, E. Sosinski, J.-F. Soussana, E. Swaine, N. Swenson, K. Thompson, P. Thornton, M. Waldram, E. Weiher, M. White, S.J. Wright, S. Zaehle, A.E. Zanne and C. Wirth, "TRY – a global database of plant traits". Global Change Biology, vol. 17, June 2011, pp. 2905-2935, doi: 10.1111/j.1365-2486.2011.02451.x

[21] S. Koch and M. Hagglund, "Health informatics and the delivery of care to older people". Maturitas, vol. 63(3), July 2009, pp.195-199.

[22] M.A. Laporte, I. Mougenot and E. Garnier, "ThesauForm – Traits: a web based collaborative tool to develop a thesaurus for plant functional diversity research". Ecological Informatics, vol. 11, September 2012, pp. 34-44, doi:10.1016/j.ecoinf.2012.04.004

[23] M.A. Laporte, I. Mougenot and E. Garnier, U. Stahl, L. Maicher and J. Kattge, "A semantic web faceted search system for facilitating building of biodiversity and ecosystems services". In: H. Galhardas & E. Rahm [Eds], Data Integration in the Life Sciences, DILS 2014, pp. 50-57. Springer, Switzerland, doi: 10.1007/978-3-319-08590-6_5

[24] J.D. Little, "Models and Managers: The Concept of a Decision Calculus". Management Science, vol. 16(8), April 1970, B-466-B485.

[25] J.S. Madin, S. Bowers, M.P. Schildhauer and M.B. Jones, "Advancing ecological research with ontologies". Trends in Ecology and Evolution, vol. 23(3), March 2008, pp. 159-168.

[26] G. M. Marakas, Decision support systems in the twenty-first century, 1999, Prentice Hall, Inc. Upper Saddle River, N.J., ISBN:0-13-744186-X

[27] M. Marschollek, "Decision support at home (DS@ HOME)–system architectures and requirements." BMC medical informatics and decision making, May 2012, 12/43, 8 pp., doi: 10.1186/1472-6947-12-43.

[28] W.K. Michener and M.B. Jones, "Ecoinformatics: supporting ecology as a data-intensive science". Trends in Ecology & Evolution, vol. 27(2), February 2012, pp. 85-93, doi: 10.1016/ j.tree.2011.11.016

[29] J. Misyak, C. Giupponi and P. Rosato, "Towards the development of a decision support system for water resource management". Environmental Modelling and Software, vol. 20(2), February 2005, pp. 203-214, doi: 10.1016/j.envsoft.2003.12.019

[30] D. Nute, W.D. Potter, F. Maier, J. Wang, M. Twery, H.M. Rauscher, P. Knopp, S. Thomasma, M. Dass, H. Uchiyama and A. Glende, "NED-2: an agent-based decision support system for forest ecosystem management". Environmental Modelling & Software, vol. 19(9), September 2004, pp. 831-843, doi: 10.1016/ j.envsoft.2003.03.002.

[31] M. Oprea, M. Sanchez-Marré and F. Wotawa, "A case study of knowledge modelling in an air pollution control decision support system". AI Communications, Binding Environmental Sciences and AI, vol. 18(4), December 2005, pp. 293-303, ISSN:0921-7126

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

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

[34] D.P.C. Peters, K.M. Havstad, J. Cushing, C. Tweedie, O. Fuentes, and N. Villanueva-Rosales, "Harnessing the power of big data: infusing the scientific method with machine learning to transform ecology". Ecosphere, vol. 5(6), Art. 67, June 2014, 15 pp., http://dx.doi.org/10.1890/ ES13-00359.1

[35] M. Poch, J. Comas, I. R-Roda, M. Sànchez-Marrè and U. Cortés, "Designing and building real environmental decision support", Systems Environmental Modelling & Software, vol. 19(9), September 2004, pp. 857-873, doi: 10.1016/j.envsoft.2003.03.007

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

[37] V. Rajasekaram and K.D.W. Nandalal "Decision Support system for Reservoir Water management conflict resolution". Journal of water resources planning and management, vol. 131(6), November 2005, pp. 1-10, doi: 10.1061/(ASCE)0733-9496(2005)131:6(410).

[38] O.J. Reichman, M.B. Jones and M.P. Schildhauer, "Challenges and opportunities of open data in ecology". Science, vol. 331(6018), February 2011, pp. 703-705, doi: 10.1126/science.1197962

[39] M. Sànchez-Marrè and K. Gibert, "Improving ontological knowledge with reinforcement in recommending the data mining method for real problems". In Proceedings of Conferencia de la Asociación Española para la Inteligencia Artificial (CAEPIA), Albacete, 9-12 November 2015, (*in press*).

[40] S.D. Cardoso,· F.K. Amanqui, K.J.A. Serique, J.L.C. dos Santos, D.A. Moreira, "SWI: A Semantic Web Interactive Gazetteer to support Linked Open Data". Future Generation Computer Systems, vol. 54, January 2016, pp. 389-398, doi: 10.1016/j.future.2015. 05.006

[41] U. Varanon, C.W. Chan and P. Tontiwachwuthikul, "Artificial Intelligence for monitoring and supervisory control of process systems". Engineering applications of artificial intelligence, vol. 20(2), March 2007, pp. 115-131, doi: 10.1016/j.engappai.2006. 07.002

[42] R.L. Walls, J. Deck, R. Guralnick, S. Baskauf, R. Beaman, S. Blum, S. Bowers, P.L. Buttigieg, N. Davies, D. Endresen, M.A. Gandolfo, R. Hanner, A. Janning, L. Krishtalka, A. Matsunaga, P. Midford, N. Morrison, É.Ó. Tuama, M. Schildhauer, B. Smith, B.J. Stucky, A. Thomer, J. Wieczorek, J. Whitacre and J. Wooley, "Semantics in support of biodiversity knowledge discovery: An introduction to the biological collections ontology and related ontologies". PLoS ONE, vol. 9(3), e89606, March 2014, doi: 10.1371/journal.pone.0089606

[43] J. Wieczorek, D. Bloom, R. Guralnick, S. Blum, M. Döring, R. Giovanni, T. Robertson and D.Vieglais, "Darwin Core: An evolving community-developed biodiversity data standard". PLoS ONE, vol. 7(1), e29715, January 2012, doi:10.1371/journal.pone. 0029715.