Knowledge Discovery for Biodiversity: from Data Mining to Sign Management

Noel Conruyt¹, David Grosser¹, Régine Vignes Lebbe²

¹LIM - IREMIA – EA 25-25 – Laboratory of Mathematics and Computer Science
University of Reunion Island, 2 rue Joseph Wetzell, 97490, Sainte-Clotilde, France
²LIS, UMR 7207-CR2P, Université Pierre et Marie Curie, Paris Universitas, France
Noel.Conruyt@univ-reunion.fr

Abstract: Knowledge discovery from data in environmental sciences is becoming more and more important nowadays because of the deluge of information found in databases of digital ecosystems, coming altogether from institutions and amateurs. For example in biodiversity science, all these data need to be validated by specialists with the help of Intelligent Environmental Decision Support Systems (IEDSSs), then enhanced and certified into qualitative knowledge before reaching their audience. Data mining through classification or clustering is the dedicated inductive process of grouping descriptions based on similarity measures, then building classes and naming them. Later, the formed concepts can be reused for identification purpose with new observations. The problem is that when using such knowledge-based systems, we tend to forget the fundamental role of subjects (end-users) in the definition, observation and description of objects. In order to get good identification results, a consensus must be found between these experts and amateurs for interpreting correctly the observed objects. Thus, a new method of Knowledge discovery is necessary by switching from Data mining to Sign management. The method focuses on the process of building knowledge by sharing signs and significations (Semiotic Web), more than on knowledge transmission with intelligent object representations (Semantic Web). Sign management is the shift of paradigm for Biodiversity Informatics that we have investigated in such domains as enhancing natural heritage with ICT. In this paper, we will present Sign management and illustrate this concept with two knowledge bases built in La Reunion Island for corals' classification with IKBS (Iterative Knowledge Base System) and plants identification with Xper2 software platform.

Keywords: Data mining; Sign management; Knowledge discovery; Environmental sciences; Biodiversity informatics.

1 KNOWLEDGE DISCOVERY IN ENVIRONMENTAL SCIENCES

At the end of the 20th century, a lot of researches have been made in computer science for enhancing the use of computers for data analysis (Breiman et al. 1984) (Diday 1994), machine learning (Mitchell 1997) and knowledge management (Skyrme 1998). Across a wide variety of fields such as science, business, industry, data have been collected and accumulated at a dramatic pace, generating a new challenge for the 21st century, called the Big Data problem (Nature 2008) (Cukier 2010). In the domain of environmental sciences, the management of natural resources such as soil, water, air, biodiversity, alimentation is facing the same deluge of information in databases (Fukuda 2009).

To tackle this problem, Fayyad et al. (1996) explained that there is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information or knowledge from the rapidly growing volumes of
digital data. These topics were the subject of the emerging field of Knowledge Discovery in Databases (KDD). KDD is the “non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. But in the digital era of the 21st century, with these torrents of new data on the Web, discovering knowledge in environmental sciences is adding a more complex challenge (Poch et al. 2004). Complexity comes from the fact that environmental data are natural (i.e. not invented by humans). They are polymorph (variable), incomplete (missing), noisy (difficult to observe), and therefore source of errors (Gilbert et al. 2008). This is particularly true for data management in the new field of Biodiversity Informatics (Sarkar 2009). It leads to different interpretations of data by people, called information1. These pieces of information can be managed to discover knowledge2 but information issued from human sensors (sensitive organs such as the eye) are more subjective than data issued from machine sensors (probes).

Thus, Knowledge Discovery in Biodiversity Informatics (KDBI) is not only a problem of extracting knowledge from a flow of data with rational mining techniques and algorithms (Gilbert 2010). It is also a problem of human and machine interaction with qualitative contents. Quality is a matter of good interpretation, visualisation and signification of the triptych (data, information and knowledge) from a subject viewpoint. Indeed, the challenge of KDBI is to make sense of the deluge of contents and forms of delivery altogether in space (geo location of data), in time (cycling or iterative process of generating knowledge) and at a certain level of observation (multi-layered study’s scale).

In fact, KDBI is embedded in the Life Science problem of understanding emerging theories (Umwelt, Activity and Semiotics) and tools that are the foundations of the concept of Sign management (Uexküll 1926) (Engeström 1987) (Peirce 1965). For taking the right decision (i.e. make a diagnosis or identify a problem) and act properly, the application of computer-assisted tools to extract knowledge from experts in knowledge bases (by representation and classification of objects) is not enough; Ontology building and case-based reasoning with the description logics formats (OWL, RDFs, SPARQL) of the Semantic Web brings syntactical advantages, but this is not sufficient for its utility and usability. We do need to know the purpose of knowledge discovery and formulate the right question before acquiring data. The alternative that we suggest for specimens’ identification is then to apply a more natural or human-centred approach to knowledge discovery by applying descriptive logics of the Semiotic Web, based on the aim to tackle real usage. Our idea is to manage not only objective knowledge with formal representations for ontologies (written notations). We must also share subjective know-how or interpretations of knowledge with living significations (multimedia annotations) that can be shown by individual subjects. As suggested by Frey (2008), let us reinvent a more living knowledge discovery on the Web for biodiversity management in the 21st century!

2 BIODIVERSITY INFORMATICS

Biodiversity Informatics is a young and rapidly growing field that brings information science and technologies to bear on the data and information generated by the study of organisms, their genes, and their interactions. In doing so, it is creating unprecedented global access to information on biological species and their role in nature (e-Biosphere 2009). The term “Biodiversity Informatics” is generally used in the broad sense of applying to computerized handling of any biodiversity information and must not be confused with the term “Bioinformatics” that

1 A Datum is a raw element of Information that is attached to a thing or entity (Object) and qualifies it with a Value measured by an instrument.
2 Information is the interpretation of a Datum measured on an Object by a Subject. It is the shaping of the datum in a message that circulates between Subjects by a channel. The interpreter makes a more or less explicit characterization of the Datum with an Attribute (character or variable) to qualify the Object.
3 Knowledge is the result of negotiating pieces of information from different Subjects by analyzing and synthesizing them (through comparison, discrimination and generalization).
complex specimens. At the same time, live memory capacities of computerizes data in the specialized area of molecular biology. At University of Reunion Island, our application domain in environmental sciences has been to manage Biodiversity with Information and Communication Technologies (ICT) at the level of Systematics. It is the scientific discipline that deals with listing, describing, naming, classifying and identifying living beings (Matile et al. 1987) (Winston 1999). In this domain, specific IEDSSs are needed to help taxonomists describe data for classification and identification purposes (Conruyt and Grosser 2007) (Ung et al. 2010). This phenetic domain of knowledge discovery, also called alpha e-taxonomy or cyber-taxonomy (Mayo et al. 2008) is concerned with ontogenetic descriptions of specimens and taxa and is not to confuse with phylogeny that deals with species history. Indeed, being able to describe, classify and identify a specimen from morphological characters is a first step for monitoring biodiversity, because it gives access to information relative to its species name (Biology, Geography, Ecology, Taxonomy, Bibliography, Photography, etc.). These tasks can be assisted in Biodiversity Informatics with a databases management system for storing unstructured information, and a knowledge bases management system for acquiring codified data and knowledge (Conruyt et al. 2008). These modules are part of a Biodiversity Information System that we want to open to other platforms such as GBIF4 and VIBRANT5 e-infrastructures by using Web Services (Conruyt et al. 2010).

3 METHOD OF KNOWLEDGE DISCOVERY

In Grosser et al. (2010), we present an inductive approach for identifying specimen descriptions with domain knowledge. An expert is assisted with computer science decision support tools for helping him to define a descriptive model (ontology), describe related cases, classify these descriptions and let other biologists identify specimens. For these complex tasks, classical discrimination methods developed in the frame of data analysis or machine learning, such as classification (Breiman 1984) or decision trees (Quinlan 1993) or others methods developed in the data mining field such as association rules mining (Platetsky-Shapiro 1991), Multifactor dimensionality reduction (Zhu and Davidson 2007) or in computer vision (Mansur et al. 2006) are not sufficient. They do not cope with relations between attributes, or with missing data, and they are not also very tolerant to errors in descriptions. The considered problem that we are faced with is to determine the class of a structured description that is partially answered and eventually contains errors, from a referenced case base, this last one being a priori classified by qualified experts in k-classes. The proposed discrimination method proceeds by inference of successive neighbouring. It is inductive, interactive, iterative and semi-directed. It combines inductive techniques of discriminatory variables and neighbours search, with the help of a similarity measure that takes into account the structure (dependencies of variables) and the content (missing and unknown values) of descriptions. It is based on the description space (inferior half-lattice) to guarantee the coherency of descriptions and the computation of generalized descriptions. Finally, the aim of our symbolic and numeric biological objects analysis is to facilitate description and classification (for specialists) and identification (for practitioners) of complex biological specimens on the Internet.

3.1 From data mining for classification and identification of taxa ...

Statistical work in taxonomy has already been done for these processing tasks since forty years with computer science tools for data analysis of phenetic attribute-value pairs (Sneath and Sokal 1973), i.e. flat attributes. All these numerical approaches led to knowledge discovery from data mining in large databases (Pankhurst 1991), (Fayyad et al. 1996). On the knowledge acquisition process, artificial intelligence techniques (Frames) and object oriented languages (Smalltalk, Java) brought also symbolic representational capacities for modelling more complex specimen descriptions. At the same time, live memory capacities of

4 http://www.gbif.org, visited on 05/10/2012
5 http://www.vbrant.eu, visited on 05/10/2012
computers has grown fast to store more and more complex object descriptions. For biological objects, Lebbe (1995) mentioned that there was a real need for implementing representation models of the class-instance type, as well as following knowledge acquisition methodologies, in order to process, refine and communicate such evolving knowledge. As a result, we use Object-Attribute-Value representations that give nowadays experts and knowledge engineers the possibility to model morphological descriptions by structuring them in knowledge bases rather than in databases. Knowledge bases are applications of Knowledge Based Management Systems (KBMS) funded on a more qualitative reasoning than databases: they focus on the quality of descriptions with a lot of dependant objects, attributes and values rather than on the quantity of records (for statistical significance) with independent variables. The difference comes also from the fact that this type of knowledge bases enhances the individual level of description (collection specimens) represented by individual cases rather than the conceptual level of domain definition (taxa) represented in a record of a database or a line of a data table.

The whole qualitative work that we have done in this context of specimen description is to capture this morphological and individual complexity in an Iterative Knowledge Base System (Conruyt and Grosser 2007). In fact, acquiring structured specimen descriptions is particularly important in zoology because it gives sense to the representation of entities in description trees with body subparts, viewpoints, specializations of objects and values (figure 1).

![Figure 1. A part of description tree of a specimen of coral (Stylophora subseriata)](image)

This anthropomorphic and individual descriptive approach is less used in botanical modelling of specimens because plants are mostly considered as species entities representing populations with independent morphological characters (Operational Taxonomic Units) (Dallwitz et al. 1995). This approach at the species level is covered by another type of KBMS called Xper², which is a management system for storage, editing, analysis and on-line distribution of descriptive data (Ung et al. 2010). A description made with Xper² is represented by a set of character-states describing one species (figure 2), whereas a description made with IKBS is represented by a set of Object-Attribute-Value describing one specimen (figure 1). These two knowledge acquisition processes are complementary: they have been used on both corals⁶ and plants⁷ classification and identification tasks at University

---

⁶ [http://coraux.univ-reunion.fr/, visited on 05/10/2012]
of Reunion Island. Xper\textsuperscript{2} knowledge bases are built in intension (conceptual description), starting from species description found in monographs and applying a top-down approach for identification. IKBS knowledge bases are made in extension (class description), for enhancing the collections of specimens found in museums and applying a bottom-up approach for classification, then identification.

**Figure 2.** A description list of a species of plant (*Trochetia granulata*).

The first method is interesting when experts want to transmit an encyclopaedic knowledge rather quickly and when taxa are well known and stable. Some difficulties reveals once the domain knowledge evolves with new knowledge (a new specimen description, a new character), and that consequently the application needs to maintain its consistency (Pullan et al. 2000). The second method is more time consuming for the expert who has to consider the intra-specific variation of species and decide how many specimen descriptions he will store in the knowledge base.

### 3.2 ... to sign management for definition and description of specimens

For our coral species identification process, specimen entities to describe are complex: they are altogether individuals, objects and populations. Colonies that are found in museum collections are composed of corallites (objects) that form the habitat of polyps (individuals). The colony (population of polyps) presents polymorphic characters with conjunction of values (simultaneous presence of states or values). It expresses intra-specific variability that depends on the biological and geographical sampling context of colonies (reef zone, lightning, profoundness, waves, etc.). Le Renard and Conruyt (1994) have specifically designed descriptive logics for making descriptive models of these biological entities. These descriptive logics of the Semiotic Web must not be confused with the description logics of the Semantic Web. Description logics (OWL, RDF) are the propositions of computer scientists that conceive syntactical formal representations for expressing ontologies and descriptions on the business object side of the machine’s interaction. Conversely, descriptive logics are propositions of domain experts for guiding the building of ontologies that are meaningful and understandable by the community. These other significant logics are user-centred; they constitute the usage object side of the interaction with humans. First, it makes sense that a description tree is better suited for visualization with dependent

\[7 \text{[http://mahots.univ-reunion.fr/, visited on 05/10/2012]}\]
objects and attributes (the absence of components causes the absence of sub-components and the irrelevance of its attributes, see figure 1). Second, the knowledge acquisition of these descriptions is made at a multi-level scale with macroscopic and microscopic observations; these viewpoints are sources of error (noisy characters, what to observe, influence of light under the binocular and the microscope) and can lead to a lot of unknown and missing values. So, it is important to establish visually the spatial level of observation by introducing these viewpoints in the descriptive model (ecology, geography, morphology, molecular, etc.). The structuration of the domain could also propose the temporal level of observation of a specimen by introducing life cycles as another dimension of the specimen description (i.e. larval, juvenile, adult). Third, modelling such ontologies closes the world of possible descriptions. The expert defines what is observable in the domain at the root of the description tree. All observed descriptions are then comparable by using the same frame of reference. He can decide to update his descriptive model by changing objects, attributes and values and then iterate on the description process. Moreover, different persons can give their own description of the same specimen in the case base. This was verified as a source of robustness for identification purpose to avoid misinterpretations of end-users (Conruyt and Grosser 2007). The design of a descriptive model with the help of descriptive logics and the possibility to introduce inter-observer variation is what makes sense of creating useful and usable knowledge bases, then used because the service espouses the needs of end-users for teaching and learning observational sciences. This communication process is called signification or construction of the Sign. It is at the root of a new paradigm for utility, usability and effective use of e-services in context called Sign management.

A Sign (figure 3) is the interpretation of an Object communicated by a Subject at a given time and place, which takes into account its content (Data, fact, event), its form (Information), and its meaning (Knowledge).

Sign management involves end-users with researchers and entrepreneurs for making them participate to the design of the product/service that they want. The problem that we have to face when making knowledge bases is that their usefulness depends on the correct interpretation of questions that are proposed by the system to obtain a good result of identification. Hence, in order to get a right identification, it is necessary to acquire qualitative descriptions. But these descriptions rely themselves on the observation guide that is proposed by the descriptive model. Moreover, the definition of this ontology is dependent upon easy visualization of descriptive logics. At last, the objects that are part of the descriptive model must be explained in a thesaurus for them to be correctly interpreted by targeted end-users. Behind each Object, there is a Subject that models this Object and gives it an interpretation. In life sciences, these objects can

![Figure 3. What is a Sign?](image-url)
be shown to other interpreters and this communication between Subjects is compulsory for sharing interpretations, and not only transmitting knowledge. The challenge of Sign management for KDBI is to involve all types of end-users in the co-design of Sign bases for them to be really used (e-service). It is why we emphasize the instantiation of a Living Lab in Teaching and Learning at University of Reunion Island (Conruyt 2010) for sharing interpretations of objects and specimens on the table rather than concepts and taxa in the head of subjects.

VALIDATION

Our knowledge management methodology has been applied in the domains of Mascarene Islands corals classification and plants identification for several years now. For our experts on corals of the Mascarene archipelago, the iterative process of knowledge acquisition is the most important task for building a robust knowledge base. IKBS brings them a reflexive tool for evaluating their observed descriptions in regards to the evolution of their projections in observable descriptive models. For families of coral knowledge bases, there are more than 30 iterations with IKBS before the expert finds a robust ontology that well explains his interpretations of observations. In Conruyt and Grosser (1999), we analysed the results of the identification process with end-users. We have found that better identification results are produced when we introduce other non-expert descriptions of the same specimens in the initial case base. Consequently, the expert, aware of the importance of transmitting his knowledge to other biologists, postulates more precise and relevant characters that may be easier to observe and/or offer less ambiguous values (easier to interpret) in his descriptive model. For example, he will refine on the basis of mutually exclusive values, monosemic attributes, frames of reference, warning signals, and enhanced illustrations on noisy questions. Misinterpretations between end-users are also at the root of identification problems for plant identification. As we can watch it in the film on the BACOMAR project on an endemic family of the Mascarene Islands (Dombeyoideae), a lot of illustrations are needed to explain the terminology and characters to observe when using Xper2. This is the reason why it is mandatory to pass from knowledge transmission with data mining to know-how sharing with sign management. A collaborative workspace for teaching and learning will be the physical and virtual space for exchanging interpretations of objects between the different actors that want to manage biodiversity.

CONCLUSION

Since several years, we proposed a knowledge management method based on top-down transmission of experts' knowledge, i.e. acquisition of a descriptive model and structured cases and then processing of these specimens' descriptions with decision trees and case-based reasoning. We designed tools to build knowledge bases. But the fact is that Knowledge is transmitted with text, not shared with multimedia, and there is a gap between interpretations of specialists and end-users that prevents these lasts from getting the right identification. Today, we prefer to deliver a sign management method for teaching and learning how to identify these specimens on a Co-Design or Creativity Platform. This bottom-up approach is more pragmatic and user-centred than the previous one because it implicates end-users at will and is open to questions and answers. The role of experts is to show amateurs how to observe, interpret and describe specimens in situ and on the table. The responsibility of biomaticians (biologists and computer scientists) is to store and share experts' interpretations of their observation, i.e. know-how rather than knowledge in sign bases with multimedia annotations for helping experts to define terms, model their domain, and allow end-users to describe and identify correctly the objects.

---

8 [http://coraux.univ-reunion.fr/spip.php?article101&lang=en, visited on 05/10/2012]
9 [http://www.canal-u.tv/video/universite_de_la_reunion_sun/bacomar_mahots.7598]
ACKNOWLEDGMENTS

We are grateful to the PO-FEDER 2-06 ICT measure for supporting our NEXTIC and BACOMAR projects: [http://www.reunioneurope.org/UE_PO-FEDER.asp] and to FP7 VIBRANT project [www.vibrant.eu] for giving us access to the International Biodiversity Informatics community for enhancing new ideas and concepts.

REFERENCES


Fukuda, K., Computer-enhanced knowledge discovery in environmental science, Phd of the University of Canterbury, 2009.


Grosser, D., Conruyt, N., and Ralambondrainy, H., Identification with iterative nearest neighbors using domain knowledge, Proceedings of the International
Sarkar, I.N., Biodiversity Informatics: the emergence of a field, BMC Bioinformatics, 10(14), 2009. 
Skyrme, D.J., Measuring The Value Of Knowledge, Business Intelligence, 1998. 