Biological and Geographical application tool

Miguel Pereira

Abstract

Monitoring and species identify is an essential step to natural resources management. Among these, biological resources aim special concern because data availability is highly limited by a number of sampling logistic constraints and catalog. As a matter of fact, data availability is one of important limitation to knowledge and, as a consequence the development of a new concept in natural resources. Data organization in digital format is a common practice in our days, and has the power to contribute to timely decisions process. Digital support and software tools are increasingly required in knowledge base systems. With this article it is our aim to present a Biological and Geographical application tool (BioGeoDB). We intend to show the application potentialities uses, the developed model and articulation between subsets. The developed application use a biological database associated with a desktop map environment. Application generates datasets reports and thematic or modeling cartographic maps display. BioGeoDB has been developed with the purpose of making compatible different datasets, derived from fieldwork studies of biodiversity monitoring programs, and also published studies. Until application development data storage was been recorder into fragment system, with no updatable structure and missing information consequences.

Introduction

Digital alphanumeric datasets have been collected and organized as a fundamental and basic instrument for decision making in knowledge process. The development of databases gives us the possibility to register species sightings with confident spatial reference. This has been the aim of some works at different geographic scales, for example the Banc de Dades de Biodiversitat de Catalunya[1], Europaea Fauna[2], National Biological Information Infrastructure[3], Species 2000[4]. Databases are useful to identify gaps, to congregate and integrate data, to prevent duplications of efforts, and to give information in monitoring programs at the national and international levels.

Environmental health and legislative frame demands quality information in species biodiversity and natural resources Research institutions and Non Governmental Organizations (NGOs) accumulate substantial species records, from fieldwork or bibliographic catalogs. Strategic information is not always easy to find, and therefore the efficient management is essential to identify relevant information in the decision making process. Collecting records does not always signify accessing to the information: Why? There are many blocking reasons; most of them related with organization and quality (e.g. corrupt files) and management issues (e.g. lost files). Quality is an essential aspect of information competitiveness[5]. Recording biological species involves different issues and computer science problems: as scale collecting, temporality, heterogeneity and validity.

Point surveys (visual/trapping) and transects are main methodologies to collecting data of species locations. The used methodology must be associated with the specie group, e.g. in bird species is usual to stay in a site or cross visual transects. Mammals are usually identified by residual fingerprints, e.g. deject or landscape marks. Beetles, dragonflies or butterflies by traps. Species occurrence depends also from temporality issues, like season of the year, and day and night cycle. Organized data enable better understand of species and habitats. The registration of species allows us to assess and to quantify biodiversity at different scales (space-time), with obvious potentialities in the definition policies[6]. Long-term data series of species communities and populations improve decisions of present trends[7].

The development of informatics tools is essential to exploring biological data[8].

Databases with biological purposes cannot exclude geographical questions and therefore must allow the identification of locations for modeling and management purposes[9]. The association of databases with GIS tools increases data utilities and promotes more general intentions of current research, making possible the emergence of new interdisciplinary fields of knowledge, allowing holistic point of view.

BDBioGeo was the natural follow up of a project created in 1996, called Alentejo's BioGeographic Unit (UNIBA). This former application was a regional database. UNIBA did not respond to new demands, as consequence of constrains and limitations in spatial modeling functions. It was decided to develop a new model, with extended functions and to whole Portugal mainland. BDBioGeo was developed with the idea of data center on biological species distribution.

Objectives

Biological data in publish paper and digital support started to be available from different sources. It was an opportunity to deal with thousands of no structured records.

What was our problem - how could data modeling produce better information? Give correspondent solution to the increasing biological data collection. The solution should include a visual data management in desktop map environment and cartographic modeling.

Most of data sources provided from fieldwork studies carried out between 1995 and 2005 or still in progress, mainly in Alentejo region. Data collections were first keep in support files (usually in sheet files) to all groups of species and geo-reference with different scale accuracy and field methodologies. As a matter of fact this sheet files were disperse in different computers with no logical and metadata edition in a fragmented data bank. Fragmented data bank means functional and updated

[1] FONT *et al.*, 2004.
[2] Fauna Europaea, 2004.
[3] CAMPBELL, 2003.
[4] WHITE, 2003.
[5] MICHAEL *et al.*, 2003.
[6] MAIER *et al.*, 2001.
[7] BOWKER, 2000.
[8] NIELSEN *et al.*, 2000.
[9] SALEM, 2003.

information lost. Available data does not mean available information, because sometimes we didn't know that data exists.

We identify three main goals:

- Record fragment datasets from monitoring and biodiversity studies[10] that have not been published yet;

- Record bibliographical data (published) on the distribution of species in Portugal mainland;

- Modeling cartographic outputs of biological data with geo-reference location and habitat sources.

With the application development we expect to improve knowledge in:

- Questions related to data environmental management;
- Sampling species in Portuguese mainland (Figura 4);
- Fieldwork facilities in oriented surveys.

Design application solution

The solution to achieve our goals was found in dual architecture of software productivity integrated application, based on commercial software as Relational Database Management System (RDBMS) and desktop map (Figura 1).

In this model, applications are dedicated to tasks, the RDBMS deals with the alphanumeric data in a way independent of desktop map. This means that decisions supported answers can be providing based on data management, as it will be showed in forward examples. The connection and data transfer between the database and the desktop map was made by the Open Database Connectivity (ODBC) protocol, which is a middleware of sufficiently simplified use[11]. With this type of application model, the alphanumeric and the graphic data are linked together for information flux[12].

Conceptual database model

The database was expected to make compatible a set of flexible premises, between the direct data sources (field studies) and indirect data sources (bibliographical). Conceptual model and translated physical database start with identification of required variables and those of essential registration (Figure 2).

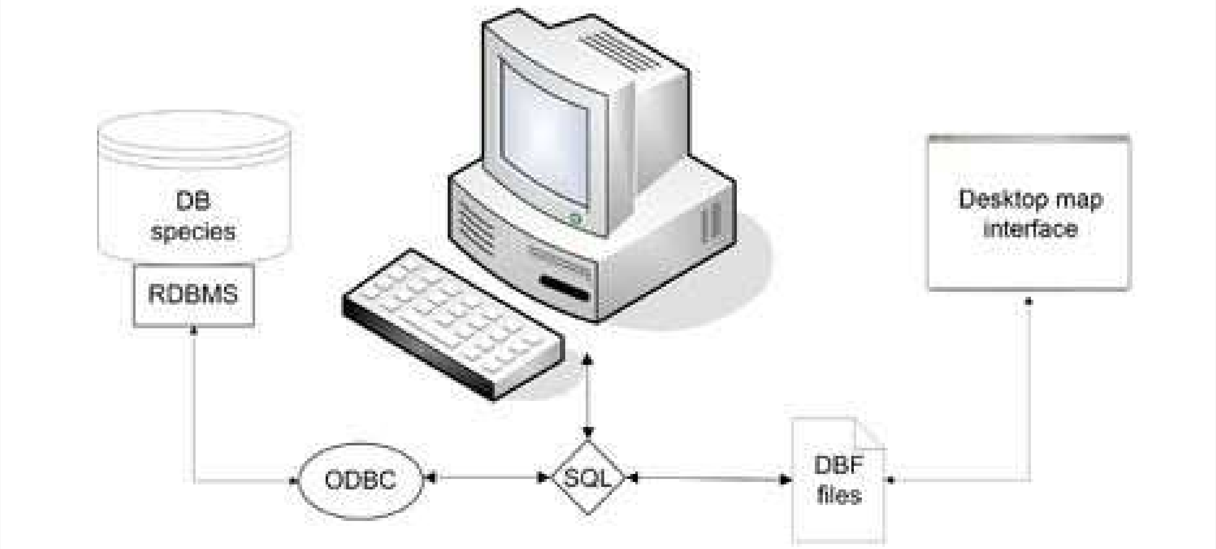


Fig.1. Dual software architecture based on RDBMS and desktop map.

[10] More in http://www.cea.uevora.pt/umc Research.
[11] ADAM & GANGOPADHYAY, 1997.
[12] PEREIRA, 2002.

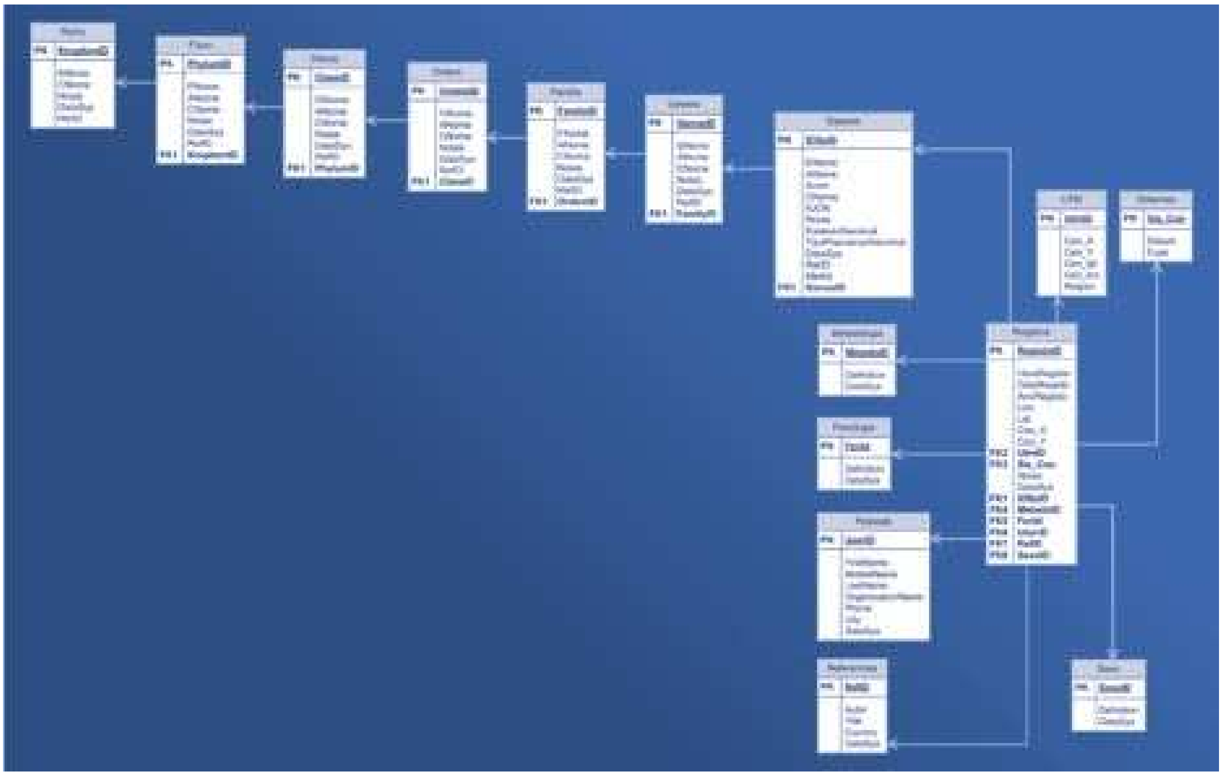


Fig.2. Alphanumeric data model, the flow between the entities reflects its Relationship.

A data collection includes the identification of the species, site location and date of the sighting observation, often the year. As we will see forward species attributes are split in tables, join information means go through several tables.

Users interact with tables; the database is structured according to a relational model[13]. This database provides analytical and generalized use principles, with specific goals structured according to the analysis needs, in a long-term perspective. Since at present it is still a restricted database, its design has been conducted taking into account analytical aspects, storage very large registry sets without major concerns to procedural questions[14].

In summary, the principles that have underlined the design of the database model, has been:

- Easiness to work (simplicity and pre-defined queries);
- Easiness of maintenance (update);
- Integration with desktop map.

Entities

The entities obey of hierarchy structure in relational model and organized as a function of species taxonomy. The model was oriented to SPECIE (Figure 2) entity level, which represents the core of the alphanumeric attributes. SPECIE entity plays a central role, connecting the relations upward (upper entities), and downward (records and relations with the geographic entity) in the "chain model". The SPECIE entity is also a checklist for the species set placed exactly before the insert records, controlling data quality. Species are registered on a vertical form, to which a unique numeric code is attributed without mismatch. Each new species is forcibly linked to a hierarchical superior taxonomic record: for example, a new species name must belong to an already existing genus. If specie belongs to a new genus, the genus should be first insert and validated. The same procedure is required for higher taxonomical levels. This rule is the result of the application of integrity law[15] that hinders the

[13] DATE, 1986; SILBERSCHATZ *et al.*, 1997.
[14] HALE & BUFFUM, 2000; Greenwald *et al.*, 2001.
[15] CODD, 1990.

existence of orphan records, i.e., without link. The SPECIE attributes also include video, photo or sounds that confer to the database visual and sensorial analysis.

The relationship between different entities is carried out through the use of primary keys in posterior entities, for example (KINGDOM.KingdomID in PHYLUM.KingdomID). This structure is based in one-to-many relationships (1:n). Records are all stored in the RECORDS entity, where species' observations including location and date. Parallel to this hierarchical structure, there are some horizontal controls and validation variables, which are all directly related with the RECORDS entity. Grouped within a set of five categories:

1. SAMPLING
- Methodogy - different species monitoring methodologies used in field survey.

2. BEHAVIOR
- Behavior - behavioral characteristics of the species, with particular focus on birds;
- Phenology - phenological characteristics of the species, with particular focus on birds.

3. CONTROL
- Refernces - data sources;
- Users - identification and the skill quantification of the researcher.

4. AMOUNT
- Abundance - number of observed individuals that is registered, using an ordinal scale.

5. GEOGRAPHIC
- Coo_Sys - geographical coordinates of point sampling records;
- UTM_Grid - Universal Transversal Mercator (UTM) 10x10 km grids, identified by unique ID's;
- Habitats - habitat description.

6. REPRODUCTION
- Age - age of individuals when sampling methods allow (applicable in mammals and birds);
- Sex - sex of individuals when sampling methods allow (applicable in mammals and birds).

The geographic identification (ID of UTM 10x10km grid) is replicated in the RECORDS entity, which makes relationship between the alphanumeric data and the spatial feature.

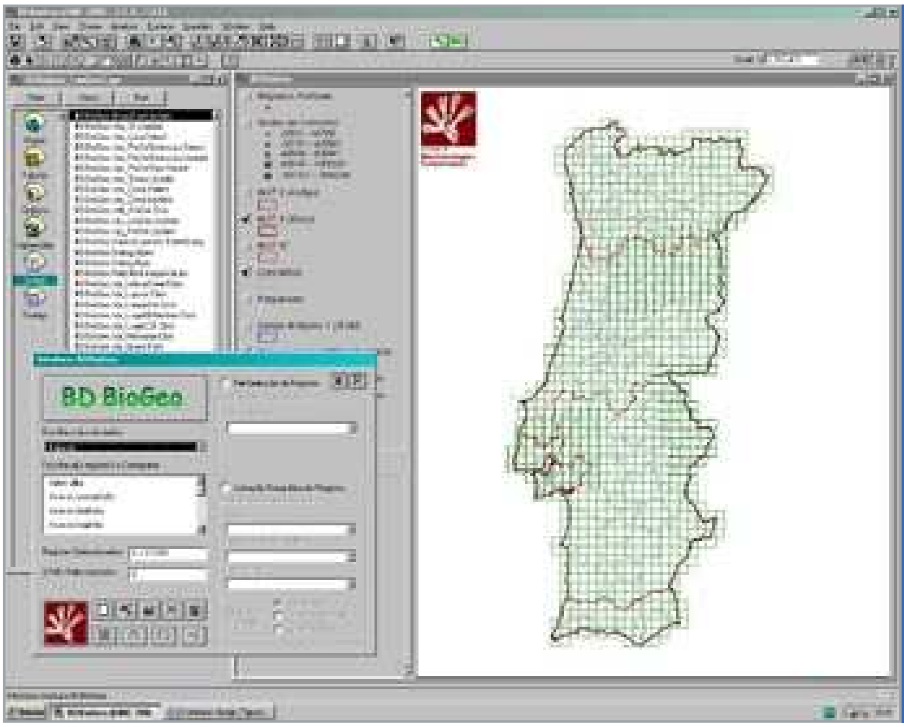


Fig.3. GIS interface menu where the users chosen species selections. Selections can be execute from different attributes (e.g. name, family, distance, administrative boundary, etc).

Desktop map frame

Desktop map interface was customized and programming based on Avenue^{TM} language (Figura 3).

The developed interface allows the user to direct query database records in different spatial or alphanumeric criteria. The automatic generated outputs allow users immediate contact with species distribution or spatial queries. Administrative boundaries or other represented features define some of must usually spatial queries. Queries can be made in different aspects of records details. The present application version allows some of regular queries and demands:

- What species present in a spatial feature?
- How many records species present in a spatial feature?
- What is the distribution of identified specie? (Figura 4)
- Neighborhood analyses (Figura 5)
- Species' richness at UTM grid or point
- Automatic reports of taxonomy

Spatial dimension of surveys tend to be detail as possible. This means that when is possible the species sightings are recorder as geographic point or planar coordinates. However, for a better efficiency in record visualization and systematization, the use of a non-projected system (geographic system) is desirable. Record coordinates sources providing from GPS units and field maps location. As mention before, species identification methodologies have a relation with spatial layers. In the less accuracy geo-reference scenario, species must at least belong to a UTM 10x10km grid. Without spatial reference is impossible to insert the specie record. This constrains as others implemented, validate and guarantee data quality and spatial location.

Implemented algorithms allow neighborhood spread functions to identify the species probability distribution (Figura 5). Based on specie distribution we sample potential distribution pattern. The values are calculated from the focal sum of near eight neighbors[16]. Habitat of species distribution is identified by and overlay theme and the record description. With the species distribution we cross other themes and new resulted information.

Results

The most obvious and useful advantage of the BioGeoDB was the centralization of all biological data in a single database. The application developed allows an easy delimitation to identify the species distribution, its occurrence or abundance at a given location, as well as the number of registered sightings.

With the centralization we make use of a fundamental tool in information handling, indispensable to researchers and institutions. Make use and manipulate information for investigation purposes; modeling species, namely their responseto different scales of sampling distributions, monitoring, resource and habitats management or ecological land planning, support decision making. Most BioGeoDB application demands have been in the academic studies, like graduated thesis or papers contribution. The temporal (records with date) and spatial (with coordinate of species position) dimensions have been crucial for the progress of such studies. This homogeneous dataset had also the virtue of allowing validation of bias species' listings, by filtering at the very moment of records input.

The BioGeoDB includes at the moment a total of 271296 registers, in a set of 2794 distinct species, observed in 996 distinct UTM grids (in 1029 possible), and 7153 distinct points. Records of terrestrial macro-invertebrates species are of special interest, since they usually absent in similar databases. As a recent article shows, mammals and birds are usually the groups of research attention[17]. BioGeoDB database includes a set of 142 species of Insecta and 99 of Arachnida. As for the comer groups, a comparison with other national and international sources of species data for Portugal mainland shows similar species representation (Table 1).

For this comparison we considered only the most commonly studied groups, since there are no references for the other groups. Some of the differences (Amphibia and Reptilia) are explained by recent changes in species taxonomy. For example, the Europaea Fauna considers some new species, not yet included in the majority of taxonomical databases.

[16] TOMLIN, 1990.
[17] KNEGTERING *et al.*, 2005.

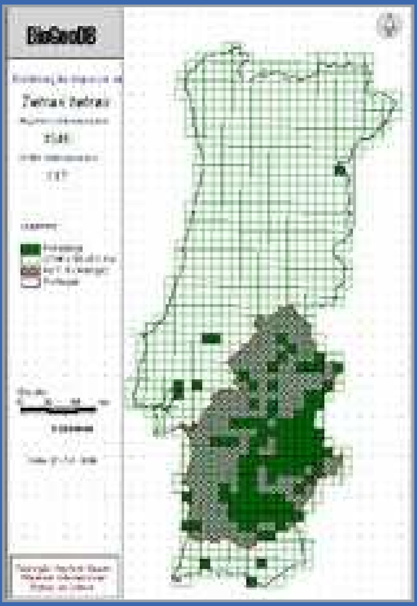
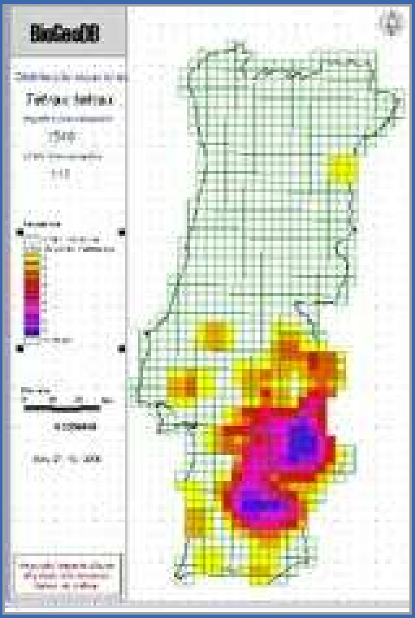


Fig.4. Distribution (solid green) of steppe birds Tetrax tetrax.

Fig.5. Neighborhood function of Tetrax tetrax distribution grids.



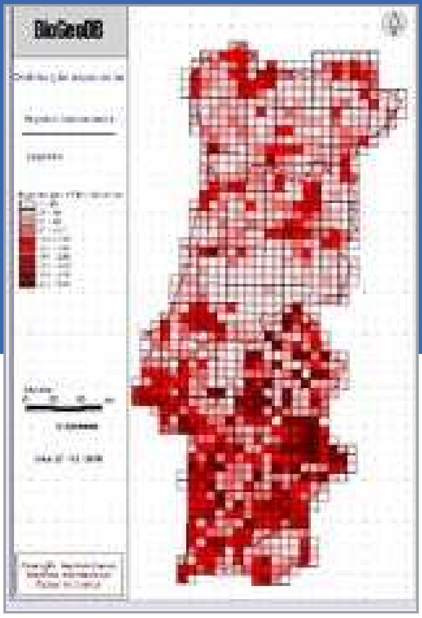


Fig.6. Count of species richness in UTM grid. The database reflects regional data sources with large number of species in Alentejo.

Conclusions of species richness and distribution should account data sources monitoring efforts; in this case we have much more data in Alentejo region (Figura 6). The cartography of species is a working instrument for the sustainable use of resources, allowing the execution of useful syntheses.

Number of species

	Amphibians	Birds	Mammals	Reptiles	Total
BioGeoDB	**18**	**357**	**78**	**29**	**482**
ICN[18]	17	274	89	27	407
Fauna Europaea	20	387	68	36	511
Iberian Fauna[19]	17	214	63	29	323

Table 1. Comparison of the main groups of species in different sources of data.

Conclusions

We appoint de simplicity and dedicated functions of the dual architecture as positive aspects. As a negative aspect we mention the retrieve process.

The environmental issues become an imperative for life quality policies. Databases will therefore increasingly be part of environmental monitoring and the security of everyday life[20]. The importance of computer science to biological knowledge is increasing; especially with spatial functions like species distribution. We have now new approaches for data analysis that otherwise would not be possible to perform, like satellite imagery instruments for spatial and temporal series. Either through remote control in local surveys, it will be necessary to make use of data to be able to act efficiently and to find fast answers during moments of crisis. With rigorous data structure it is possible to monitoring more efficiently the alterations that terrestrial ecosystems are subject to.

The potential of BioGeoDB is now starting to be explored, in analysis of temporal series by comparing historical records. We intend to make useful of application developed to perform planning decisions. BioGeoDB has potential for forthcoming projects providing sightings data The fact that a sufficiently homogeneous species dataset is provided, in comparison with other databases, allows us to state that it has enough quality for its use in biodiversity studies. In the future the decisive challenge will be to have summarized contents available on the Internet. The future goal will be therefore to create a platform that makes available its contents to the scientific community and the public, such as maps of distribution or species reports ■

Acknowledgments

My first acknowledgment goes to Pedro Roque and Rui Raimundo by them dedicated work on Avenue scripts and taxonomic issues and as a team members. To all dedicated UMC technicians and researchers by the support, especially Prof. Diogo Figueiredo. All team members of monitoring programs. The Funding support PORA, Sub-program C-Measure 2.

References

ADAM, R.N. & GANGOPADHYAY, A. (1997), *Database Issues in Geographic Information Systems*, Boston. Kluwer Academic Publishers..

[18] Instituto Conservação da Natureza, 2005.
[19] RAMOS *et al.*, 2001.
[20] Commission of the European Communities, 2005.

BOWKER, G., Work and Information Practices in the Sciences of Biodiversity, In a Presented, *Proceedings.26th International Conference on Very Large Batabases, Cairo*, Egypt. 10 (October 2000).

CAMPBELL, J., *National Biological Information Infrastructure Entrerprise Architecture*, Center for Biological Informatics of the U.S.Geological Survey. http://www.nbii.gov/about/pubs/enterprise_architecture/NBII_Design_Architecture.pdf, 2005

CODD, E.F. (1990), *The Relational model for database management; version 2*, Massachusetts, Addison Wesley.

Commission of the European Communities., *Global Monitoring for Environment and Security (GMES): Establishing a GMES capacity by 2008 - Action Plan (2004-2008)*, http://europa.eu.int/eur-lex/en/com/cnc/2004/com2004_0065en01.pdf, 2005

DATE, C.J. (1986), *An introduction to Database Systems* 4ª Ed, Rio de Janeiro, Editora Campos.

Fauna Europaea version 1.1, *Fauna Europaea Web Service*, http://www.faunaeur.org, 2004.

FONT, X., CÁCERES, M., NAVARRO, A., and QUADRADA., *Banc de Dades de Biodiversitat de Catalunya,* Generalitat de Catalunya and Universitat de Barcelona. http://biodiver.bio.ub.es/biocat/homepage.html, 2004.

GREENWALD, R., STACKOWIAK, R., and STERN, J. (2001), *Oracle Essentials : Oracle 9i & Oracle 8i & Oracle8,* Sebastopol O'Reilly & Associates, Inc.

HALE, S.S., BUFFUM, H.W. (2000), Designing Environmental Databases for Statistical Analyses, *Environmental Monitoring and Assessment*, 64:55-68.

Instituto Conservação Natureza., *Sistema de Informação do Património Natural*, http://www.icn.pt/sipnat/sipnat1.html, 2005.

MAIER, D., LANDIS, E., CUSHING, J., FRONDORF, A., SILBERSCHATZ, A., and SCHNASE. J. (2001), Research Directions in Biodiversity and Ecosytem Informatics, *NASA Workshop on Biodiversity and Ecosystem Informatics held at NASA Goddard Space Fight Center, June 22-23, 2000.*

MICHAEL, G., MILLER, K.T., KAREN, L., MARY, A.C., and JOANNA, B. (2003), An ecologically oriented database to guide remediation and reuse of contaminated sites, *Remediation Journal, 14:69.*

NIELSEN, E., JAMES, E., and MEREDITH, L. 2000, Biodiversity Informatics: The Challenge of Rapid Development, Large Databases, and complex Data, In a Presented, *Proceedings 26th International Conference on Very Large Batabases, Cairo*, Egypt. 10 (October 2000).

PEREIRA, M., (2002), Uso de DesktopMap para manipulação de informações biogeográficas em SIG, GeoFocus: *International Review of Geographical Informacion Science and Technology* (http://www.geo-focus.org), 33-48.

KNEGTERING, E., DREES, J., GEERTSEMA, P., HUITEMA, H., & UITERKAMP, A. (2005), Use of Animal Species Data in Environmental Impact Assessments, *Environmental Management,* 36:862-871.

RAMOS, M, LOBO, J.M., and ESTEBAN, M. (2001), Ten years inventorying the Iberian fauna: results and perspectives, *Biodiversity and Conservation,* 10:19-28.

SALEM, B. (2003), Application of GIS to biodiversity monitoring, *Journal of Arid Environments*, 54:91-114.

SILBERSCHATZ, A., KORTH, H.F., and SUDARSHAN, S. (1997), *Database System Concepts*. São Paulo. MAKRON Books.

TOMLIN, C.D. (1990), *Geographic Information Systems and Cartographic Modeling,* New Jersey, Prentice Hall.

WHITE, R.J., Species 2000 - *Common Data Model*, http://www.species2000.org/index.html, 2003.

Miguel Pereira, Investigador no centro de Ecologia e Ambiente da Universidade de Évora. masp@uevora.pt