Marine biological data, as many other types of data, are often collected on a relatively narrow geographical scope, and over a short time span. Yet, in order to answer questions related to global change, we need to study long-term, large-scale patterns. Thus there is a need to bring together data to enable discovery of regional, global, and long-term patterns.

The following questions are still vital today: how do we integrate individual datasets into massive databases that are able to support these studies? How can we, as a community, ensure that we are covering the whole field, and that no taxonomic groups or geographical areas are left out? How can we compare data collected by different methods?

The organisation of this effort can benefit from cost-efficiencies and synergies of effort. How can we make maximal use of resources, and avoid overlaps? What is the role of international organisations such as ICES, the IOC and FAO in this? What is the role of CoML and OBIS, and of GBIF? Which others have a role to play?

The 'Ocean Biodiversity Informatics' conference offered a forum to marine biological data managers to discuss the state of the field, and to exchange ideas on how to further develop marine biological data systems. The conference took place in Hamburg, Germany, from 29 November to 1 December 2004 and was jointly organised by the Flanders Marine Institute (VLIZ), the Intergovernmental Oceanographic Commission of UNESCO - International Oceanographic Data and Information Exchange (IOC/IODE), the International Council for the Exploration of the Sea (ICES), the Census of Marine Life - Ocean Biogeographic Information System (CoML/OBIS), the International Association for Biological Oceanography (IABO), the Taxonomic Database Working Group (TDWG), the Marine Biodiversity and Ecosystem Functioning EU Network of Excellence (MarBEF) and was hosted by the Bundesamt für Seeschifffahrt und Hydrographie (BSH). 168 delegates came from all over the world, including 37 countries, and from national, inter-governmental and non-governmental organisations, universities and data centres.

ISSN 1377-0950

IOC VLIZ BUNDESAMT FÜR SEESCHIFFFAHRT UND HYDROGRAPHIE OBIS OCEAN BIOGEOGRAPHIC INFORMATION SYSTEM CENSUS OF MARINE LIFE

Ocean Biodiversity Informatics
Proceedings 2007

IOC Workshop Report No. 202
VLIZ Special Publication No. 37

Proceedings

Ocean Biodiversity Informatics

International Conference on Marine Biodiversity Data Management
Hamburg, Germany
29 November - 1 December, 2004

Edited by
Edward Vanden Berghe,
Ward Appeltans,
Mark J. Costello,
and Peter Pissierssens

IOC Workshop Report No. 202
VLIZ Special Publication No. 37

PROCEEDINGS

OCEAN BIODIVERSITY INFORMATICS

International Conference on Marine Biodiversity Data Management

Hamburg, Germany
29 November to 1 December 2004

Edited by

Edward Vanden Berghe[1], Ward Appeltans[1], Mark J. Costello[2] and Peter Pissierssens[3]

[1]Flanders Marine Institute (VLIZ), Wandelaarkaai 7, B-8400 Oostende, Belgium
[2]University of Auckland; Leigh Marine Laboratory, Box 349, Warkworth, New Zealand
[3]United Nations Educational, Scientific and Cultural Organization (UNESCO); Intergovernmental Oceanographic Commission (IOC), 1 rue Miollis, F-75732 Paris Cedex 15, France

This symposium was jointly organised by the Flanders Marine Institute (VLIZ), the Intergovernmental Oceanographic Commission of UNESCO - International Oceanographic Data and Information Exchange (IOC/IODE), the International Council for the Exploration of the Sea (ICES), the Census of Marine Life - Ocean Biogeographic Information System (CoML/OBIS), the International Association for Biological Oceanography (IABO), the Taxonomic Database Working Group (TDWG), the Marine Biodiversity and Ecosystem Functioning EU Network of Excellence (MarBEF) and was hosted by the Bundesamt für Seeschifffahrt und Hydrographie (BSH).

IOC Workshop Report No. 202
VLIZ Special Publication No. 37

Disclaimer

Bibliographic reference

Vanden Berghe, E., W. Appeltans, M.J. Costello, P. Pissierssens (Eds). Proceedings of 'Ocean Biodiversity Informatics': an international conference on marine biodiversity data management Hamburg, Germany, 29 November - 1 December, 2004. Paris, UNESCO/IOC, VLIZ, BSH, 2007. vi + 192 pp. (IOC Workshop Report, 202) (VLIZ Special Publication, 37)

Picture cover

F. Nuyttens

Co-published in 2007 by

United Nations Educational, Scientific and Cultural Organization
Intergovernmental Oceanographic Commission
1 rue Miollis
F-75732 Paris Cedex 15
France

Vlaams Instituut voor de Zee vzw
Flanders Marine Institute
Wandelaarkaai, 7
B-8400 Oostende
Belgium

Bundesamt für Seeschiffahrt und Hydrographie
Bernard Nochtstrasse 78
Postfach 301220
D-20305 Hamburg
Germany

Printed in UNESCO's workshops

Printed in France

(SC-2007/WS/1)
ISSN 1377-0950

Preface

We know very little about the biodiversity in the world's oceans. But one thing is sure: the diversity of the type of data and information that is stored in data systems around the world is increasing dramatically. In two important meetings, the first in Hamburg in 1996, the second in Brussels in 2002, biologists have discussed how to take an example from the physical oceanographers, and to formulate plans on how to work together to integrate individual databases. Developments in technology have made possible new approaches to data sharing and dissemination. Distributed databases are becoming a reality, and the advantages of a distributed system now far outweigh the extra cost of technical complexities to create them.

The International conference on Marine Biodiversity Data management 'Ocean Biodiversity Informatics' was held in Hamburg, Germany, from 29 November to 1 December 2004. Its objective was to offer a forum to marine biological data managers to discuss the state of the field, and to exchange ideas on how to further develop marine biological data systems. Many marine biologists are actively gathering knowledge, as they have been doing for a long time. What is new is that many of these scientists are willing to share their knowledge, including basic data, with others over the Internet. Our challenge now is to try and manage this trend, avoid confusing users with a multitude of contradicting sources of information, and make sure different data systems can be and are effectively integrated.

Acknowledgements

Our sincere thanks go to our sponsors, who made this conference possible, and to the scientific community that made organizing this symposium such a pleasure: the European Union, World Data Center for Oceanography A in Silver Spring, US, and the German Federal Ministry for the Environment, Nature Conservation and Nuclear Safety.

To attend the conference MarBEF members received support from the MarBEF Network of Excellence 'Marine Biodiversity and Ecosystem Functioning' which is funded by the Sustainable Development, Global Change and Ecosystems Programme of the European Community's Sixth Framework Programme (contract no. GOCE-CT-2003-505446). This publication is contribution number MPS-07005 of MarBEF.

Thanks also to Daphne Cuvelier, Simon Claus and Leen Vandepitte for their comments and corrections on some of the papers and Maren Fabricius and Ward Appeltans for their help with the administrational and practical organization of the symposium itself.

Table of contents

Ocean Biodiversity Informatics – an emerging field of science

Edward Vanden Berghe[1] and Mark J. Costello[2]

[1]Flanders Marine Institute (VLIZ), Wandelaarkaai 7, B-8400 Oostende (Belgium)
wardvdb@vliz.be

[2]University of Auckland; Leigh Marine Laboratory, Box 349, Warkworth (New Zealand)

Abstract

'Ocean Biodiversity Informatics' (OBI) heralds a new era in biological research and management that is revolutionising the way we approach marine biodiversity research. OBI uses computer technology to manage marine biodiversity information (capturing, storing, searching for, retrieving, visualising, mapping, modelling, analysing and publishing data). This allows more users better and faster access to biodiversity information than ever before. The global nature of phenomena such as climate change, over-fishing, and other changes in ecosystems, would not have been recognised had it not been for informatics-aided analyses.

The prospect of data mining and exploration on a global scale is enough to gladden the hearts of marine scientists across the world, as marine biology embraces the computer age. Access to global data through OBI will allow for worldwide gap analysis resulting in new perspectives on current research, the promotion of collaborations between research groups and real data sets for teaching purposes, to mention just a few of the potential benefits. OBI is an initiative of the 21st century and will make conventional marine biodiversity research more dynamic and comprehensive, with a range of constantly evolving online tools.

Background

We know very little about the biodiversity in the world's oceans. But one thing is sure: the diversity of the type of data and information that is stored in data systems around the world is increasing dramatically. While well-managed databases with global coverage used to be restricted to geophysical sciences, this is no longer true. In two important meetings, the first in Hamburg in 1996, the second in Brussels in 2002, biologists have discussed how to take an example from the physical oceanographers, and to formulate plans on how to work together to integrate individual databases. In the workshop held in Hamburg in 1996, discussions were held on how to improve the quantity and quality of chemical and biological data available to the scientific community. The specific purpose of the workshop was to provide recommendations to guide management of chemical and biological oceanographic data by the Programme on International Oceanographic Data and Information Exchange (IODE) of the Intergovernmental Oceanographic Commission (IOC).

Biological data managers met with their physical oceanography colleagues during the 'Colour of Ocean Data' symposium held in Brussels, 25-27 November 2002. While it was realised there that the needs of biological data managers were different from those of physical oceanographers, it was stressed that commonalities are more important than differences. Some applications were presented that demonstrated the power of collaboration across disciplines.

Data on marine species and their environment (be it regional or on a global scale) is the 'fuel' on which OBI operates and therefore complements the traditional disciplines of taxonomy, ecology and biogeography. However, many scientists are ignoring pleas from international scientific organisations, including the International Council for Science (ICSU), to make their data public. The irony is that most data collections are paid for, directly or indirectly, by public funds. Taxonomists have led the way with regard to public accessibility of data, as type specimens are lodged in museums for the common good. Molecular sequence data is to be submitted to international repositories, where the data is publicly available, before research on these sequences is accepted for publication in scientific journals. It is suggested that there should be a protocol for scientists where ecological data would be made available in a similar way. OBI will make more data available to more people more quickly than ever before, including the repatriation of data and information collected in developing countries.

A change in biological science culture to one of open access to primary data is essential for accountability of research. A greater recognition of the value of biodiversity data accessibility by the scientific community, publishers, funding agencies and employers is vital. This change in culture is currently underway. Many marine biologists are actively gathering knowledge, as they have been doing for a long time. What is new is that many of these scientists are willing to share their knowledge, including basic data, with others over the Internet. The World Wide Web makes it easy and cheap to share data in a global community. It is compared to the invention of printing centuries before, and like it, will bring more information and opportunities for discovery to people all over the world. Our challenge now is to try and manage this trend, avoid confusing users with a multitude of contradicting sources of information, and make sure different data systems can be and are effectively integrated.

Meta-databases and other methods of data discovery will certainly gain more and more importance as the number of studies, and the number of scientists conducting these studies, increases. Such methods of data discovery, and better communication between and among scientists and data managers, are essential for avoiding unnecessary duplication of effort. Comprehensive inventories of collected data should facilitate projects of data archaeology and rescue, and make sure knowledge about the existence of data does not disappear.

One of the most challenging aspects of any informatics system is quality assurance. This is particularly important when you consider all the possible uses to which data can be put, and the number of steps from the original point of data collection to the end use.

However a data or information system is designed, its continuity and development depend on support from the scientific community. This community includes contributors, evaluators of funding applications, users and science policy makers. While not long ago it was difficult to find any information at all, now the Internet user is confronted with a large number of possible sources, but without an indication of the quality of the proffered information. Again, metadata can play a role in this; it should contain information on who collected the data, and for which purpose. The first can be an indication of quality, or lack thereof. The latter can be an indication of 'fitness for use', *i.e.* an evaluation of whether the data would be suitable for a particular analysis.

Integrating data from different sources brings its own set of problems. One very important aspect is that one has to make sure that the same terms are used in the different component datasets – for taxonomy, geography and measured parameters. Several initiatives exist to compile lists of taxonomic names – some for specific taxa, or for a restricted geographical area, some global in geographic and taxonomic scope. In order to facilitate data integration, we need a world register of marine organisms, containing all taxonomic names, including recent synonyms and often-used misspellings. Such a register, of which the content should be managed by the taxonomic community, would allow 'translating' between different sets of names now in use for different data sets.

Developments in technology have made possible new approaches to data sharing and dissemination. Distributed databases are becoming a reality, and the advantages of a distributed system now far outweigh the extra cost of technical complexities to create them. The Census of Marine Life (CoML), and its data management programme Ocean Biogeographic Information System (OBIS), is making this happen for marine biogeography. The Ocean Biogeographic Information System (www.iobis.org) is already publishing almost 10 million location records for 61,000 marine species from a global network of over 100 databases. OBIS is the information component of the Census of Marine Life, the largest ever global marine biology discovery programme (www.coml.org).

The main challenge is now to remain in control of developments, and not to let OBI become technology-driven. Obviously, recent technical developments should be monitored, and implemented when they bring a significant improvement, either in efficiency or in functionality. But what we do is much more important than how we do it. The real issues for data management are standardisation, collaboration and enabling knowledge-based decision-making.

OBI, the conference

The conference topics were restricted to marine biological data management – taxon-based, biogeography but also environmental, non-taxon based data management. Specific objectives were to

- Learn how and why researchers have used large-scale marine biodiversity databases to make major discoveries about the functioning and state of ocean ecosystems.

- Bring together biological data managers to discuss the present state, and progress, in this field since the meetings in Hamburg (1996) and Brussels (2002).
- Discuss standards and protocols for data exchange. Take note of new developments such as Distributed Generic Information Retrieval (DiGIR) and OBIS, and discuss how this will influence biological data management in general.
- Provide an opportunity for biological data managers to find out what is happening at IODE National Oceanographic Data Centres and marine research agencies from around the world.
- Discuss potential gaps and overlaps in the taxonomic and geographic scope of existing data systems. How can we, as a community, ensure that we are covering the whole field, and that no taxonomic groups are left behind? How can we make maximal use of resources, and avoid overlaps?
- How do we integrate data from separate databases into large datasets that will enable us to provide answers on the global cover and long time scales that we need?

The conference was organised as a series of five consecutive thematic sessions, with a final session dedicated to a panel discussion. Each of these thematic sessions had a corresponding poster session. In an opening session, several acknowledged experts were asked to review the field. Themes for the sessions were

- Opening session
- Information system development
- Taxon-based systems:
- Geography-based systems
- Analysis
- Panel discussion:

    - What is our target audience, and how effectively are we reaching it?
    - How do we integrate individual databases into datasets that allow large-scale, long-term analyses? What is the role of international organisations such as the International Council for the Exploration of the Sea (ICES), the IOC and the Food and Agriculture Organization of the United Nations (FAO) in this? What is the role of CoML and OBIS, and of GBIF? Which others have a role to play?
    - How do we avoid overlap and duplication of effort? How do we avoid gaps in taxonomic/geographical coverage?
    - What mechanisms are we using now, or do we plan to use in the future, to disseminate data? What should we do to persuade data providers to make their data available?

All sessions had ten oral presentations, except for the opening session with three scientific presentations, and the session on Analysis with 11 speakers – a total of 44 oral scientific presentations. In parallel, there was a permanent poster sessions, with a total of 56 poster presentations. A full list of titles of all presentations and posters is given on page 13.

During the opening session, participants were welcomed by Prof. Dr Peter Ehlers, President of the Federal Maritime and Hydrographic Agency (BSH), Hamburg/ Rostock. In his address, Dr Ehlers stressed the importance of integration of information over different disciplines; we have to optimise marine data management, not only biological data, but tough a multi-disciplinary approach. Mrs H. Imhoff welcomed the participants on behalf of the German Federal Ministry of the Environment; it was thanks to the financial support of her Ministry, and her own personal interest in this matter, that we were able to use the nice conference facilities. Mrs Julie Gillin, Data Manager of ICES, and Mrs Lesley Rickards, Chair of the International Oceanographic Data and Information Exchange programme of the Intergovernmental Oceanographic Commission (IOC/IODE) addressed the meeting and stressed the importance of this conference for their respective organisations. Dr Mark J. Costello and Dr Edward Vanden Berghe reviewed the objectives of the meeting, and discussed expected outcomes.

The second part of the opening session consisted of three scientific talks, giving a broad overview of different aspects of OBI. Arthur Chapman gave a review of issues of quality control and the principles of data quality; he discussed the concept of ‘Fitness for Use’, meaning that data perfectly acceptable for one purpose might be useless for another. Donald Hobern of the GBIF Secretariat talked about new and existing technologies for data integration using distributed systems – the basis of emerging global systems for biodiversity information. The third review was from Dr Daniel Pauly, on fisheries impact on global marine biodiversity and ecosystems. In his talk, Dr Pauly illustrated how large, integrated datasets can bring insights that can’t be learned from individual datasets.

The main point of discussion during the Panel Discussion was about the advantages of data sharing, and mechanisms to stimulate data custodians to make their data openly available. A draft text had been circulated from the beginning of the conference, and was edited during this final panel session. The resulting text was approved by all participants of the meeting, and is reproduced in this volume.

There were 168 participants in the conference, coming from 39 different countries; the complete list can be found on page 8. Not unexpectedly, best represented was Germany with 38 participants. This, together with the generous support from the German Government and the host institution, clearly demonstrate that OBI is taken serious in Germany. The US and UK were runner-up with 19 and 18 participants, respectively. Financial support from the EU made it possible to invite people from developing countries, and brought us participants from India, Mauritius, Seychelles, Tunisia and the Philippines. Extra support from the World Data Centre for Oceanography, Silver Spring, allowed us to support participants from South America.

Several papers, based on presentations or posters of the conference, were published as a Theme Section of the Marine Ecology Progress Series (Vol. 316, 2006). The table of contents of the Theme Section is given on page 18. Thanks to financial support from MarBEF, all papers in this Theme section are Open Access, and can be freely downloaded from the internet, either from the site of MEPS (http://www.int-res.com/abstracts/meps/v316/), or from the MarBEF web site (http://www.marbef.org/modules.php?name=moa&lvl=Ref&section=1&refid=100285).

All conference presentations, publications and other documents remain available on the site of the Flanders Marine Institute (VLIZ)(http://www.vliz.be/obi).

Organisers

Local host: Bundesamt für Seeschifffahrt und Hydrographie (BSH, Germany)
Organising committee: Maren Fabricius (BSH), Edward Vanden Berghe (IOC/IODE and VLIZ), Peter Pissierssens (IOC/IODE), Mark J. Costello (OBIS), Friedrich Nast (BSH)
Scientific Committee: Mark J. Costello (OBIS), Fred Grassle (CoML), Syd Levitus (World Data Centre for Oceanography, Silver Springs), Peter Pissierssens (IOC/IODE), Tony Rees (OBIS), Lesley Rickards (British Oceanographic Data Centre, Chair IODE), Edward Vanden Berghe (VLIZ, Chair IODE GE-BICH), Sunhild Wilhelms.

Ocean Biodiversity Informatics conference statement Hamburg 1 December 2004[1]

We note that increased availability and sharing of data:

- is good scientific practice and necessary for advancement of science,
- enables greater understanding through more data being available from different places and times,
- improves quality control due to better data organisation, and discovery of errors during analysis,
- secures data from loss.

The advantages of free and open data sharing have been determining factors while developing the data exchange policy of the Intergovernmental Oceanographic Commission of UNESCO.

We call on scientists, politicians, funding agencies and the community to be proactive in recognising data's:

- overall cost/benefit,
- importance to science,
- long-term benefits to society and the environment,
- increased value by being publicly available.

We also call upon employers of scientists, academic institutions and funding agencies and editors of scientific journals, to:

- promote on-line availability of data used in published papers,
- promote comprehensive documentation of data, including metadata and information on the quality of the data,
- reward on-line publication of peer reviewed electronic publications and on-line databases in the same way conventional paper publications are rewarded in the hiring and promotion of scientists,
- encourage and support scientists to share currently unavailable data by placing it in the public domain in accordance with publicly available standards, or in formats compatible with other users.

[1] A draft of this statement was first circulated at the beginning of the conference. This draft was presented and commented upon by all the conference participants during the last session, and during a final consultation through e-mail.

List of participants

Achuthankutty, Chittur Thelakkat: National Institute of Oceanography, India
Addink, Wouter: Universiteit van Amsterdam, Netherlands
Amaral Zettler, Linda: Marine Biological Laboratory, USA
Appeltans, Ward: Flanders Marine Institute (VLIZ), Belgium
Arvanitidis, Christos: Hellenic Centre of Marine Research, Greece
Bahri, Tarub: FAO - Medsudmed, Italy
Bailly, Nicolas: Muséum National d'Histoire Naturelle, France
Bamber, Roger: Natural History Museum, UK
Barnard, Michaela: University of Hull, UK
Bartley, Jeremy: University of Kansas, USA
Beard, Doug: U. S. Geological Survey, USA
Belov, Sergey: RIHMI.WDC RUNODC, Russia
Bertocci, Iacopo: Università di Pisa, Italy
Bijoux, Jude: Seychelles Centre for Marine Research and Technology, Seychelles
Blight, Clinton: University of Saint Andrews, Sea Mammal Research Unit, UK
Bluhm, Bodil: University of Alaska Fairbanks, USA
Boedeker, Dieter: Bundesamt für Naturschutz, Germany
Boethling, Maria: Federal Maritime and Hydrographic Agency, Germany
Branton, Robert M.: Marine Fish Division, USA
Brenner, Jorge: Technical University of Catalonia, Spain
Bründel, Jörg: Federal Maritime and Hydrographic Agency, Germany
Carbonnière, Aurélien: European Science Foundation, France
Chapman, Arthur: Australian Biodiversity Information Services, Australia
Chartosia, Niki: Aristotle University of Thessaloniki, Greece
Che-Bohnenstengel, Anne: Federal Maritime and Hydrographic Agency, Germany
Chinman, Richard: OBIS, USA
Coman, Claudia: National Institute for Marine Research and Development "Grigore Antipa" (NIMRD), Romania
Cooper, Keith: Centre for Environment, UK
Costello, Mark: University of Auckland, New Zealand
Cousineau, Patrice: Department of Fisheries and Oceans, Canada
Dadic, Vlado: Institute of Oceanography and Fisheries, Croatia
Daly Yahia, Mohamed Néjib: University of 7 November at Carthage, Tunesia
Danis, Bruno: Royal Belgian Institute of Natural Sciences, Belgium
Dargie, James: Countryside Council for Wales Headquarters, UK
D'Auria, Giuseppe: Universidad Miguel Hernández, Spain
Davies, Jon: Joint Nature Conservation Committee, UK
De Broyer, Claude: Royal Belgian Institute of Natural Sciences, Belgium
de Bruin, Taco: Royal Netherlands Institute for Sea Research, Netherlands
de Kluijver, Mario: Universiteit van Amsterdam, Netherlands
De Pablo, Maria Jesus: Universidad Autónoma de Madrid, Spain
Deprez, Tim: Universiteit Gent, Belgium
Duckworth, Kim: Ministry of Fisheries, New Zealand

Dunn, Tim: Joint Nature Conservation Committee, UK
Ehlers, Peter: Federal Maritime and Hydrographic Agency, Germany
Fabri, Marie-Claire: Institut Français de Recherche pour l'Exploitation de la Mer, France
Fabricius, Maren: Federal Maritime and Hydrographic Agency, Germany
Faulwetter, Sarah: Hellenic Centre for Marine Research, Greece
Fautin, Daphne G.: University of Kansas, USA
Fernández, Mercedes: University of Valencia, Spain
Finney, Kim: National Oceans Office, Australia
Froese, Rainer: Christian-Albrechts-University Kiel, Germany
Fyrberg, Lotta: Swedish Meteorological and Hydrological Institute, Sweden
Galparsoro, Ibon: AZTI Foundation, Spain
Garneria, David: University of Valencia, Spain
Gelfeld, Robert: National Oceanic and Atmospheric Administration, USA
Gertman, Isaac: Israel Marine Data Center, Israel
Gillin, Julie: International Council for the Exploration of the Sea, Denmark
Glöckner, Frank Oliver: Max Planck Institute for Marine Microbiology, Germany
Goodin, Kathy: NatureServe, USA
Gotsis, Panagiotis: Hellenic Centre for Marine Research, Greece
Gradinger, Rolf: University of Alaska Fairbanks, USA
Grassle, Frederick: Rutgers University, USA
Greve, Wulf: Senckenbergische Naturforschende Gesellschaft, Germany
Groß, Onno: Sciene & Journalism, Germany
Hahn, Andrea: Botanischer Garten und Botanisches Museum Berlin-Dahlem, Germany
Hällfors, Heidi: Finnish Institute of Marine Research, Suomi/Finland
Halpin, Patrick N.: Duke University, USA
Haroun Trabaue, Ricardo: University of Las Palmas de Gran Canaria, Spain
Hennessy, Martina: Marine Institute Headquarters, Ireland
Hering, Ingelore: Federal Maritime and Hydrographic Agency, Germany
Hilmi, Karim: Institut National de Recherche Halieutique, Morocco
Hobern, Donald: Global Biodiversity Information Facility, Denmark
Hoeksema, Bert: National Museum of Natural History - Naturalis, Netherlands
Huang, Jr-Chuan: Academia Sinica, Taiwan
Hughes, Michael: British Oceanographic Data Centre, UK
Huguet, Antoine: Institut Français de Recherche pour l'Exploitation de la Mer, France
Imhoff, Heike: German Federal Ministry for the Environment, Germany
Ivankovic, Damir: Institute of Oceanography and Fisheries, Croatia
Jahncke, Kai: Bundesamt für Seeschifffahrt und Hydrographie, Germany
Kameya, Albertina: Instituto del Mar del Peru, Perú
Kaschner, Kristin: University of British Columbia, Canada
Kedra, Monika: Polish Academy of Sciences, Poland
Kemp, Zarine: University of Kent Computing Laboratory, UK
Kiefer, Dale: System Science Applications, USA
Kirmitzoglou, Ioannis: Aristotle University of Thessaloniki, Greece
Kitsos, Miltiadis-Spyridon: Aristotle University of Thessaloniki, Greece
Klenz, Birgitt: Federal Research Centre for Fisheries, Germany
Koester, Burkhard: Senckenbergische Naturforschende Gesellschaft, Germany
Kohnke, Dieter: n/a, Germany
Kozyr, Alexander: Carbon Dioxide Information Analysis Center, USA

Kraberg, Alexandra: Alfred Wegener Institut für Polar- und Meeresforschung, Germany
Kullander, Sven: Swedish Museum of Natural History, Sweden
Künitzer, Anita: The Federal Environmental Agency, Germany
Lakkis, Sami: Lebanon University, Lebanon
Larsen, Lena: International Council for the Exploration of the Sea, Denmark
Latuhihin, Max: Ministerie van Verkeer en Waterstaat, Netherlands
Lehfeldt, Rainer: Federal Waterways Research and Engineering Institute, Germany
Lin, Jack: Academia Sinica, Taiwan
Lion, Monica: Ministerio de Ciencia y Tecnología, Spain
Lleonart, Jordi: Food and Agricultural Organisation of the United Nations, Italy
Lombardot, Thierry: Max Planck Institute for Marine Microbiology, Germany
Lowry, Roy: British Oceanographic Data Centre, UK
Lüerßen, Gerold: Common Wadden Sea Secretariat, Germany
Lyakh, Anton: Ukraine National Academy of Sciences, Ukraina
Lykiardopoulos, Aggelos: Hellenic Centre for Marine Research, Greece
Maggi, Elena: Università di Pisa, Italy
Maillard, Catherine: SISMER, France
Martinez Arbizu, Pedro: Senckenbergische Naturforschende Gesellschaft, Germany
Meerhaeghe, Angelino: Royal Belgian Institute of Natural Sciences, Belgium
Miranda, Ana: Ministerio de Ciencia y Tecnología, Spain
Moncheva, Snejana: Bulgarian Academy of Sciences, Bulgaria
Moncoiffé, Gwenaëlle: British Oceanographic Data Centre, UK
Motamedi, Khosro: Federal Maritime and Hydrographic Agency, Germany
Moura, Gisela: FH - OOW, Germany
Myroshnychenko, Volodymyr: Middle East Technical University, Turkey
Nast, Friedrich: Federal Maritime and Hydrographic Agency, Germany
Neal, Phillip: Marine Biological Laboratory, USA
Nedderhut, Holger: N/A, Germany
O'Brien, Todd: National Oceanic and Atmospheric Administration, USA
Palacios, Carmen: Max Planck Institute for Marine Microbiology, Germany
Palazov, Atanas: Bulgarian Academy of Sciences, Bulgaria
Panov, Vadim: Russian Academy of Sciences, Russia
Parr, Jon: Marine Biological Association, UK
Paterson, Gordon: Natural History Museum, UK
Patterson, Edward J.K.: Suganthi Devadason Marine Research Institute, India
Pauly, Daniel: University of British Columbia, Canada
Pissierssens, Peter: UNESCO, France
Pool, Wim: Royal Netherlands Institute for Sea Research, Netherlands
Poonyth, Asha: Mauritius Oceanography Institute, Mauritius
Quast, Christian: Max Planck Institute for Marine Microbiology, Germany
Rantajärvi, Eija: Finnish Institute of Marine Research, Suomi/Finland
Rees, Tony: Commonwealth Scientific & Industrial Research Organisation (Australia), Australia
Ricard, Daniel: Dalhousie University, Canada
Richter, Michael: Max Planck Institute for Marine Microbiology, Germany
Rickards, Lesley: British Oceanographic Data Centre, UK
Rius, Josephine France: Worldfish Center, Philippines
Rodriguez-Valera, Francisco: Universidad Miguel Hernández, Spain

Rojas, Ricardo: Hydrographic and Oceanographic, Chile
Romero, Arturo: Oceanography Institute of the Ecuadorian Navy, Ecuador
Rosenberg, Gary: Academy of Natural Sciences, USA
Rühl, Niels-Peter: Federal Maritime and Hydrographic Agency, Germany
Sagen, Helge: Institute of Marine Research, Norway
Schaap, Dick: Marine Information Service, Netherlands
Schratzberger, Michaela: Centre for Environment, UK
Schwabe, Reinhard: Federal Maritime and Hydrographic Agency, Germany
Sergeyeva, Oleksandra: Ukraine National Academy of Sciences, Ukraina
Shiganova, Tamara: P.P. Shirshov Institute of Oceanology RAS, Russia
Smirnov, Igor S.: Russian Academy of Sciences, Russia
Stevens, Darren: The Sir Alister Hardy Foundation for Ocean Science, UK
Stocks, Karen: University of California, USA
Szaron, Jan: Swedish Meteorological and Hydrological Institute, Sweden
Taconet, Marc: Food and Agricultural Organisation of the United Nations, Italy
Teeling, Hanno: Max Planck Institute for Marine Microbiology, Germany
van Soest, Rob: Universiteit van Amsterdam, Netherlands
van Spronsen, Edwin: Universiteit van Amsterdam, Netherlands
Vanden Berghe, Edward: Flanders Marine Institute (VLIZ), Belgium
Vanhoorne, Bart: Flanders Marine Institute (VLIZ), Belgium
Vladymyrov, Vladimir: UNESCO, Belgium
von Dorrien, Christian: Federal Research Centre for Fisheries, Germany
Wafar, Mohideen: National Institute of Oceanography, India
Warzocha, Jan: Sea Fisheries Institute, Poland
Webb, Andy: Joint Nature Conservation Committee, UK
Wesnigk, Johanna: Max Planck Institute for Marine Microbiology, Germany
Wilkinson, Steve: Joint Nature Conservation Committee, UK
Zhang, Yunqing: Rutgers University, USA
Zimmer, Ferdinand: German Federal Ministry for the Environment, Germany
Zimmermann, Christopher: Federal Research Centre for Fisheries, Germany
Zodiatis, George: Oceanography Center (DFMR), Cyprus

List of presentations

OPENING SESSION

A. Chapman. From data to uncertainty - principles of data quality.
D. Hobern. Architecture and standards for Global Biodiversity Informatics - a GBIF and TDWG perspective.
D. Pauly, R. Watson. Fisheries impact on global marine biodiversity and ecosystems: inference from large heterogeneous data sets.

INFORMATION SYSTEM DEVELOPMENT

B.D. Best, P.N. Halpin. Emerging open source software, standards and protocols used for sharing and analysing marine biogeographic data.
K. Duckworth. The application of standardised data quality improvement methodologies to data describing marine biodiversity: three case studies illustrating the New Zealand experience.
K. Finney. Key ingredients for developing a national oceans portal.
A. Hahn, A. Kirchhoff, W.G. Berendsohn. Networking biodiversity data – online access to distributed datasources in GBIF-D.
R. Lehfeldt. *The* role of metadata in the management of data from the coastal zone.
R. Lowry, L. Bird, P. Haaring. A semantic modelling approach to biological parameter interoperability.
T. Rees. Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System.
D. Schaap. Pan-European network for ocean and marine data & information management – (SEA-SEARCH).
I. Shevchenko, G. Moiseenko, O. Vasik. Metadata exchange with use of ontologies.
K.I. Stocks, K.S. Baker. How do you build an information system that works? Lessons from environmental case studies.

TAXON-BASED SYSTEM

T. Deprez, M. Vincx, E. Vanden Berghe, J. Mees. NeMys: an evolving biological information system, a state of art.
H. Enevoldsen, M. Lion, C. Sexto, B. Sims. The IOC-ICES-PICES harmful algal event data-base, HAE-DAT.
R. Froese. Lessons learned in the design, data management and user needs of FishBase.
W. Greve. Communicating the properties of marine organisms as the second dimension of marine biodiversity.
S. Lavery, H. Ross, M. Dalebout, M. Goode, G. Ewing, P. McLaren, A. Rodrigo, C.S. Baker. Validating marine biodiversity: interconnecting web-based DNA taxonomy and geography.
J. Lleonart, M. Lamboeuf, M. Taconet. FAO databases on marine species identification for fishery purposes.

I.S. Smirnov, E.P. Voronina, A.L. Lobanov, A.V. Neyelov. The information system of the marine animals collection (fish and invertebrates) in the Zoological Institute of the Russian Academy of Sciences.
S.G. Lelekov, Yu.N. Tokarev, V.V. Melnikov, N.P. Pakhorukov, A.M. Lyakh, E.Yu. Georgieva, S.A. Tzarin, V.F. Zhuk, Yu.B. Belogurova, O.N. Danilova. TAXEX: TAXonomical EXpert System - history of development and technology of identification.
E. Vanden Berghe, P. Bouchet, G. Boxshall, M.J. Costello, C. Emblow. European Register for Marine Species version 2.0: data management, current status and plans for the future.

GEOGRAPHY-BASED SYSTEM

C. Arvanitidis, V.D. Valavanis. MedOBIS: biogeographic information system for the Mediterranean and Black Sea.
M. J. Costello, F. Grassle, Y. Zhang, K. Stocks, T. Rees. The evolution and challenges of the ocean biogeographic information system.
J. Davies, S. Wilkinson , D. Connor. Managing and distributing marine biodiversity data to meet the needs of marine conservation.
M.D. Fortes. Marine biodiversity conservation in Asia-Pacific: are scientific data effectively managed?
P. Halpin, A. Read, L. Crowder; B. Best, D. Hyrenbach, S. Freeman. OBIS-SEAMAP: developing a biogeographic research data commons for the conservation of marine mammals, sea birds and sea turtles.
D.A. Kiefer, F.J. O'Brien, M.L. Domeier. A new environmental information system for tracking tagged marine organisms.
L.I. Larsen. Trawl survey data stored at ICES.
S. Levitus, R. Gelfeld. Building ocean profile-plankton database: progress since the first "International workshop on oceanographic biological and chemical data management".
T.D. O'Brien. Building a global Plankton database: eight years after Hamburg 1996.
D. Stevens, A. Richardson. The continuous Plankton recorder database: current uses and future directions.

ANALYSIS

R. Branton, D. Ricard. Using OBIS to provide reliable regional-scale estimates of population indices for marine species from research trawl surveys.
J. Brenner, J.A. Jimémez. Spatial database model of ichthyofauna bioindicators of coastal environmental quality.
M.J. Costello, C.S. Emblow, P. Bouchet, A. Legakis. An analysis of gaps in knowledge of marine biodiversity in Europe.
P. Cousineau, J.R. Keeley. Integrating marine information and products using the Canadian Geospatial Data Infrastructure (CGDI).
V. Dadic, D. Ivankovic. MEDAS System for archiving, visualisation and validation of oceanographic data.
D. G. Fautin, J. M. Guinotte, B.A. Maxwell, J. D. Bartley, A. Iqbal, R. W. Buddemeier. Predicting and understanding biogeographic ranges from occurrence

records and correlated environmental data: a method-development study using clownfishes and their sea anemone hosts.
A. Kozyr, R.M. Key. Global Ocean Data Analysis Project(GLODAP): results and data presentation through CDIAC data visualisation tools.
D. Ricard, R.A. Myers, L. Lucifora, F.Ferretti, A. Porter. Integrating the OBIS schema into the information system that supports the Pew Global Shark Assessment.
G. Rosenberg. Species naming curves versus species accumulation curves.
W. Zhuang, Y. Zhang , J.F. Grassle. Identify erroneous distribution data in OBIS using outlier detection techniques.

POSTERS

W. Addink, M. Brugman. Achieving data integration and solving interoperability problems when connecting databases to distributed systems.
A. Lyakh, I. Agarkova-Lyakh. BioFOX Zoo: A software for storing and analysing zooplankton sampling data.
J.C. Alba-Casado, R. Pushker, F. Rodríguez-Valera. Micro-Mar: a marine prokaryotes database to correlate habitat with taxonomy.
W. Appeltans, E. Vanden Berghe and J. Mees. A taxonomic and biogeographic information system of marine species in the southern North Sea developed by Flanders Marine Institute.
A. Aramburu, A. Borja, I. Galparsoro, Y. Sagarminaga. A marine observatory network for the Basque Country (n.e. of Spain).
K. Bradtke, L. Szymanek, A. Krężel. Thermal fronts in the Baltic Sea and their influence on spatial distribution of ocean colour phenomena.
R.M. Branton, L. Van Guelpen. A next step in the emergence of self-funded OBIS regional nodes: industry sponsored data product development on the CMB-BIO internet portal.
V. Chavan, M.V.M. Wafar , S. Krishnan. Biodiversity informatics and Indian Ocean: challenges and potentials.
G. Coleman. An integrated approach to managing data, graphics and reports.
S.R. Coppola, V. Giacalone , T. Bahri. Fisheries and ecosystems information system: a tool for the implementation of the ecosystem approach to mediterranean fisheries.
J. Dulčić, B. Grbec, L. Lipej, G. Beg-Paklar, N. Supić & A. Smirčić. The effect of the hemispheric climatic oscillations on the Adriatic ichthyofauna.
Z. Erdoğan, C. Turan, H.T. Koç. Stock identification of European anchovy (Engraulis encrasicolus) using morphometric and meristic characters.
M.C. Fabri, J. Galeron, G. Maudire. BIOCEAN – a new database for deep-sea benthic ecological data.
D. Garnería, M. Fernández, J.A. Raga. A Mediterranean Geo-database of cetacean strandings (MEDACES): an implement for research and conservation.
B. Grbec, J. Dulčić, M. Morović. Adriatic-possible influence of climate oscillations over the northern hemisphere.
S. Hankin, L. Dantzler, R. Cohen , F. Grassle. Data management and communications (DMAC) for the U.S. integrated ocean observing system (IOOS).

F. Hernandez, B. Vanhoorne R. T'Jampens, E. Vanden Berghe. Web mapping services for biological data.
M. Hughes, R. Lowry. The BODC taxonomic browser – a powerful search tool to locate taxonomic information.
A. Kameya, M. Quiñe, E. Delgado, S. Sánchez, A. Chipollini. Marine biodiversity data in Peru.
K. Kaschner, R. Watson, A.W. Trites, D. Pauly. Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model.
A. Kraberg, F. Buchholz, K. Wiltshire. Dealing with the challenges in presenting taxonomic information online: an introduction to PLANKTON*NET@AWI.
M. Kędra, J. Urbański. Linear referencing as a tool for analyses of organic material deposition along a sandy beach of Gdańsk – Sopot – Gdynia (Polish coast of Baltic Sea).
B. Klenz. What information on biodiversity can be derived from ichthyoplankton surveys in the western Baltic Sea?
A.Koukouras, M.S Kitsos, I. Kirmitzoglou, C. Arvanitidis. Development of an updating information system on Decapoda Crustacea Museum collections, useful in education and scientific research.
S. Lakkis. Plankton dataset from Lebanese seawater.
J. Larsen. DOME – database on oceanography and marine ecosystems.
A. Lyakh, I. Agarkova-Lyakh. Technology of distribution phytoplankton geometric information - Using database to study the taxonomic distribution of organisms inhabiting macrophytes of the Black Sea.
A. Meerhaeghe, K.D. Cauwer, M. Devolder, S. Jans, S. Scory. Operational integration of biodiversity and physico-chemical data: experience at the BMDC.
N. Mikhalov, E. Vyazilov, S. Belov, S. Sukhonosov. JCOMM ETDMP Pilot Project – the prototype of the "End to End" marine data management technology - basic solutions and development status.
K. Millard, F. Hernandez, E. Vanden Berghe. Integration of biological information with other information classes - the experiences of the IOC/IODE MarineXML initiative.
A. Miranda, C. Eirín, G. Fernández. Mesozooplankton from Ría de Vigo (NW Spain) and its adjacent shelf between 1995 and 2003.
Ø. Moestrup, H. Enevoldsen. The IOC taxonomic reference list of toxic Plankton algae.
G. Moncoiffe. Marine biological data management at the British Oceanographic Data Centre.
V. Myroshnychenko, S.C. Polat Beken. MED POL Phase III database.
E.L. Nevrova, A.N. Petrov. Use database on the Black Sea benthic diatoms for study of its biodiversity.
T.D. O'Brien. NMFS-COPEPOD: An online, investigator-friendly, global plankton database.
T.S.Osadchaya, S.V. Alyomov, J.G.Wilson. Macrobenthic communities in relation to long-term oil pollution of the coastal bottom environment (Black Sea).
N.P. Pakhorukov. Fish biodiversity of raising in the World Ocean.
C. Palacios, B. Olsson, A. Boetius, P. Lebaron. New biotechnologies for sampling the ecological diversity of the oceans: the informatics challenge.
V.E. Panov, V.S. Shestakov. Towards an information system on aquatic invasive species with early warning functions for European coastal waters.
J. Parr, D. Lear. The Marine Life Information Network (MarLIN).

E.J.K. Patterson. Marine biodiversity data management and dissemination: a case study from Gulf of Mannar marine biosphere reserve, Southeast Coast of India.
A.N. Petrov, E.L. Nevrova. Evaluation of diversity and assessment of technogenic pollution impact upon taxocene structure of benthic diatoms in coastal zone of the Black Sea.
H. Rees, E. Rachor, E. Vanden Berghe. The ICES North Sea Benthos Project: objectives and data management.
J. Rius, G. Tolentino-Pablico, R. Froese, D. Pauly. Predicting trophic level for all fishes.
R.L. Rojas. Biological data management activities at CENDOC related to research cruises in the Chilean inner southern channels.
S. Sato, M. Nagao, N. Baba, Y. Tomioka. Marine biological data management at Japan Oceanographic Data Center (JODC).
O. Sergeyeva. Environmental sensitivity mapping.
K.I. Stocks, D. Tittensor, R.A. Myers. Using data systems to evaluate seamount biogeography.
E. Tamiro, C. Gitto, A. Bergamasco. A dynamic web portal for integrated coastal monitoring.
S.A. Tsarin. Biodiversity and functioning of the sound scattering layers (SSL) myctophidae taxocenosis in the Tropical Ocean.
V.M. Tsontos, D.A. Kiefer, F.J. O'Brien. A web-based fisheries oceanographic information system for the Gulf of Maine.
C. Turan. Genetic and morphologic divergence and phylogenetic relationships of Mediterranean Mullidae species (Perciformes).
B. Vanhoorne, S. Claus, D. Cuvelier, E. Vanden Berghe, J. Mees. EurOBIS: the European node of the ocean biogeographic information system.
E. Vyazilov, S. Sukhonosov. The most effective fields of use of XML language.

Marine Ecology Progress Series (MEPS) - Theme Section (Vol 316, 2006)

Costello M.J, Vanden Berghe E, Browman H.I. Introduction.
http://www.int-res.com/articles/meps_oa/m316p201.pdf
Costello M.J, Vanden Berghe E. 'Ocean biodiversity informatics': a new era in marine biology research and management.
http://www.int-res.com/articles/meps_oa/m316p203.pdf
Fabri M.C, Galéron J, Larour M, Maudire G. Combining the biocean database for deep-sea benthic data with the online Ocean Biogeographic Information System.
http://www.int-res.com/articles/meps_oa/m316p215.pdf
Arvanitidis C, Valavanis V.D, Eleftheriou A, Costello M.J, Faulwetter S, Gotsis P, Kitsos M.S, Kirmtzoglou I, Zenetos A, Petrov A, Galil B, Papageorgiou N. MedOBIS: biogeographic information system for the eastern Mediterranean and Black Sea.
http://www.int-res.com/articles/meps_oa/m316p225.pdf
Lleonart J, Taconet M, Lamboeuf M. Integrating information on marine species identification for fishery purposes
http://www.int-res.com/articles/meps_oa/m316p231.pdf
Halpin P.N, Read A.J, Best B.D, Hyrenbach K.D, Fujioka E, Coyne M.S, Crowder L.B, Freeman S.A, Spoerri C. OBIS-SEAMAP: developing a biogeographic research data commons for ecological studies on marine mammals, seabirds and sea turtles.
http://www.int-res.com/articles/meps_oa/m316p239.pdf
Stevens D, Richardson A.J, Reid P.C. Continuous Plankton Recorder Database: evolution, current uses and future directions.
http://www.int-res.com/articles/meps_oa/m316p247.pdf
Costello M.J, Bouchet P, Emblow C.S, Legakis A European marine biodiversity inventory and taxonomic resources: state of the art and gaps in knowledge.
http://www.int-res.com/articles/meps_oa/m316p257.pdf
Guinotte J.M, Bartley J.D, Iqbal A, Fautin D.G, Buddemeier R.W. Modeling habitat distribution from organism occurrences and environmental data: case study using anemonefishes and their sea anemone hosts.
http://www.int-res.com/articles/meps_oa/m316p269.pdf
Kaschner K, Watson R, Trites A.W, Pauly D. Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model.
http://www.int-res.com/articles/meps_oa/m316p285.pdf

Developing OBIS into a Tool to Provide Reliable Estimates of Population Indices for Marine Species from Research Trawl Surveys

Robert M. Branton[1] and Daniel Ricard[2]

[1]Bedford Institute of Oceanography
1 Challenger Drive, Dartmouth, Nova Scotia, Canada B2Y 4A2
email: BrantonB@mar.dfo-mpo.gc.ca

[2]Dalhousie University, Department of Biological Sciences,
1355 Oxford Street, Halifax, Nova Scotia, Canada, B3H 4R2

Abstract

Research trawl surveys from Canada's Department of Fisheries and Oceans (DFO) and the US National Oceanographic and Atmospheric Agency (NOAA) are used as a basis for developing a road-map to prepare research trawl surveys in general for public access via systems such as Ocean Biogeographic Information System (OBIS). Data quality issues associated with surveys includes validation of species names, treatment of zeros, data standardization techniques and provision of confidence limits. Suggestions to improve the OBIS system include support for summary statistics and length classes as well as addition of a gazetteer facility. The Bedford Institute of Oceanography's recently established OBIS provider service is also described.

Keywords: fisheries; trawl surveys; OBIS schema; data quality control.

Introduction

Traditional research trawl surveys (Doubleday and Rivard, 1981) are species rich (100s) with analysis focused only on commercial species (~10s). Recently other species (*e.g.* mega-invertebrates) have been added to sampling protocols thus enabling investigation of ecosystem issues (DFO, 2003). Current expectations are that the Ocean Biogeographic Information System (OBIS) will provide a basis for interoperability of these data with other scientific disciplines (Grassle, 2000). Using Canada's Department of Fisheries and Oceans (DFO) and the US National Oceanographic and Atmospheric Agency (NOAA) research trawl surveys we provide a basic road-map for preparing research trawl survey data sets for public access via systems such as OBIS. This presentation focuses on: DFO/NOAA trawl surveys, trawl survey data quality issues, and ways to improve OBIS. A description of the new OBIS provider service located at the Bedford Institute of Oceanography (BIO) is also given.

Methods

DFO and NOAA have been collecting data from standardized research trawl surveys on the east coast of North America since the 1960s and have considerable local expertise and software for preparing resource assessments from these data. The first significant effort to integrate these data for biogeographic studies was in 1995 as part of the East Coast North America Strategic Assessment Project (ECNASAP) (Brown *et al.*, 1996). Amongst other things, the ECNASAP project integrated basic survey catch data (numbers and weights) from 5 fisheries laboratory databases (Table 1), providing observations on 276 species from ~50,000 fishing sets for the period 1970-95. Although the surveys are ongoing, this dataset has not been updated since '95 and is presently only available on Compact Disk from project principals as a 300+ column flat file.

Table 1. The ECNASAP project integrated data from 5 fisheries laboratory databases for the period 1970-95.

Laboratory	Laboratory Location	OBIS Collection Code
North Atlantic Fisheries Centre	St. John`s, Newfoundland, Canada	DFO-NFLD
Maurice Lamontagne Institute	Mont-Joli, Québec, Canada	DFO-NG
Bedford Institute of Oceanography	Dartmouth, Nova Scotia, Canada	DFO-SF
Gulf Fisheries Centre	Moncton, New Brunswick, Canada	DFO-SG
Northeast Fisheries Science Center	Falmouth, Massachusetts, USA	NMFS-NEFSC

In 2002, BIO's Scotian Shelf Summer survey data was temporarily placed directly on the OBIS portal as an interim measure until a permanent OBIS provider service could be installed at BIO. Since 2002, BIO has developed a relational database version of the ECNASAP data for local research as well as for serving to OBIS and the 'Gulf of Maine Ocean Data Partnership'. These efforts are expected to provide a basis for DFO and NOAA to develop publicly accessible and near real-time links to all of their ongoing surveys. Providing public access to other than the basic survey catch data presently contained in the ECNASAP dataset will require careful attention to trawl data quality issues and to extending limits of the present OBIS schema.

Results

Following are the major issues to be considered when preparing data such as the DFO/NOAA and ECNASAP research trawl surveys for systems like OBIS.

Species list validation

OBIS uses the Species 2000 Catalogue of Life (CoL) annual checklist CD-ROM as its basis for validating species names. All OBIS providers are therefore recommended to use this same CoL checklist to find the most current scientific names and hierarchies for species contained in their databases. The CoL checklist is available online from http://www.sp2000.org/. A preliminary comparison between the ECNASAP species list and the Integrated Taxonomic Information System (ITIS) database, which is part of the CoL database, indicates that approximately 50 of 276 species names in the ECNASAP database are either obsolete or incorrectly spelled. Local taxonomists and survey staff should review discrepancies in the local lists and make corrections where appropriate. They should also note which species are difficult to identify or not routinely sampled. The taxonomic hierarchy data are particularly useful as they can be used to prepare cumulative discovery curves (Costello, 1996) for each hierarchy level. This type of analysis can help identify new species appearing as a result of protocol changes from those appearing as a result of environmental changes.

Taking care of zeros

Fisheries surveys are primarily intended to provide observations (*e.g.* numbers and weights) for species captured in the sampling gear. Absence of a species is not recorded during surveys and hence not recorded in the trawl survey database. Fisheries analysts work with well documented pre-established stock area definitions and employ database queries and analysis programs that automatically generate ZERO (0) for the missing data, as opposed to NULL values. ZERO is interpreted as the absence of a surveyed species from a trawl, and would be included in calculations of averages; a NULL value would signify that no information is available, and would be omitted from any further analysis. Survey species lists must also include these established stock area definitions (*e.g.* lists of survey strata) thus providing a clear indication of when and where fisheries analysts are interpreting the missing data as ZERO.

Adjusted v. standardized observations

The probability of a particular organism being retained in a research trawl depends on many factors, not least of which are fishing vessel and gear used. Data contributors should provide distinct survey series names (*e.g.* OBIS-Collection Code) for each unique survey vessel, sampling gear, stratification plan, and season combination. Given that good data management practices dictate that data be stored as they were recorded, observed values (*e.g.* observed individual count and weight at length, sex and maturity) given to end-users or to systems such as OBIS should be automatically adjusted by sampling ratio (*i.e.* total/sample). Adjusted numbers-at-age from sampled materials (*e.g.* otoliths and scales) should be based on stock specific age-length keys (*e.g.* proportion at age for given length). Observations from sets where gear has been damaged although containing rare organisms should not be given to users expecting adjusted results. End users should be further aware that not all fisheries laboratories routinely standardize their observed values for distance towed (*e.g.* standard/observed) or species for catchability by gear (*e.g.* proportion caught at length). Databases should include sufficient metadata to clearly indicate how the data at hand have been adjusted and standardized.

Confidence limits

The DFO and NOAA research trawl surveys discussed here all follow the same basic stratified random design. Relative indices such as ‘average per standard tow’ should include variance or standard error. Absolute estimates such as ‘total biomass’ and ‘total abundance’, if presented, should be peer reviewed and given with Internet links to citable publications (*e.g.* Canadian Science Advisory Secretariat - http://www.dfo-mpo.gc.ca/CSAS/).

Recommendations

These recommendations are for improving the OBIS system as a whole. Extending the schema by providing support for area summaries (*e.g.* number of observations and variance) at the same time as allowing for more details (*e.g.* length, parent catalog number) is intended to broaden the range of information and products conveyed to the public.

Add new keywords to existing schema concepts

- Basis of Record – stratum average, stock estimate.
- Locality – stratum, ecozone, grid square, stock area ...
- Life Stage – maturity stage, age class.

Add new schema concepts

- Number of Samples and Sampling Units in Locality.
- Length Class of Observed Individuals.
- Variances or Error Estimates for Observed Individual Count and Weight.
- Parent Catalog Number for stomach contents and parasites.

Add new schemas

- Collection metadata – descriptions of vessels, gears ...
- Gazetteer – stratum, ecozone, grid square, stock areas

Enhance end-user interface

DFO Maritimes routinely provides a variety of publicly available survey based data products (Branton and Black, 2003). The OBIS portal should consider providing a range of mapping products including collection based multi-species mapping and reporting using expanding pie symbol maps (*e.g.* multiple species on one map). Observations for multiple species should be optionally given by row or column, with missing values being given as zeros or nulls. In addition to set by set catch data, OBIS should also provide summary statistics by stratum, ecozones, etc. Methods that enable species catchability standardization should also be investigated.

Regional Scale DiGIR Sevices

The Global Biodiversity Information Facility (GBIF) and OBIS networks use the Distributed Generic Information Retrieval (DiGIR) protocol and the Darwin Core (DwC) schema including Darwin Core 2 (DwC2) and OBIS variants. DFO has established a regional-scale DiGIR server at BIO to enable near real-time posting of multiple datasets to the OBIS portal. BIO's DiGIR service sits within a specially controlled portion of the DFO firewall known as the Demilitarized Zone (DMZ) allowing limited connections with a controlled set of known partners. Inputs include small scale specialized databases (*e.g.* Atlantic Conservation Data Centre) provided via the File Transfer Protocol (FTP) and large institutional scale databases (*e.g.* DFO Maritime and North East Fisheries Science Center trawl surveys) via the Oracle SQL*net protocol. The only output from the DiGIR provider service is XML-formatted data sent to OBIS portal's global cache. Data flow into and out of the DiGIR provider service is in the form of pre-scheduled transfers, with all of the public queries handled by the OBIS portal. The OBIS portal manages all movement of data to the GBIF portal.

Conclusion

The suggested improvements provide a systematic basis for ongoing enhancement and extension of the OBIS schema and interface. Improved ability to integrate data from disparate sampling schemes would in turn provide a capacity to derive population/community indices of abundance, diversity, production, etc. around the world. Trophic cascade models using trawl survey and Continuous Plankton Recorder (CPR) data being developed for the Scotian Shelf (Choi *et al.,* 2004) could, for example, be tested in the North Sea.

References

Branton R.M. and J. Black. 2003. Summer Groundfish Survey Update for selected Scotia Fundy Groundfish Stocks. DFO Canadian Science Advisory Secretariat Research Document 2003/089. 60p.

Brown S.K., R. Mahon, K.C.T. Zwanenburg, K.R. Buja, L.W. Claflin, R.N. O´Boyle, B. Atkinson, M. Sinclair, G. Howell and M.E. Monaco. 1996. East Coast of North America Groundfish: Initial Explorations of Biogeography and Species Assemblages. NOAA, Silver Springs, MD, DFO, Dartmouth, Nova Scotia. 100+ pages.

Choi J.S., K. Frank, W.C. Leggett and K. Drinkwater. 2004. Transition to an alternate state in a continental shelf ecosystem. Can J Fish Aquat Sci 16(4):505-510.

Costello M.J., C.S. Emblow and B.E. Picton. 1996. Long term trends in the discovery of marine species new to science which occur in Britain and Ireland. J Mar Biol Ass U K 76:255-257

DFO. 2003. State of Eastern Scotian Shelf Ecosystem. DFO Canadian Science Advisory Secretariat Ecosystem Status Report 2003/004.

Doubleday W.G. and D. Rivard. 1981. Bottom trawl surveys. Can Spec Publ Fish Aquat Sci 58: 273p.

Grassle J.F. 2000. The Ocean Biogeographic Informations System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional context. Oceanography 13(3):5-9.

Spatial database model of ichthyofauna bioindicators of coastal environmental quality

Jorge Brenner and José Antonio Jimémez

Coastal Management Group, Departament d'Enginyeria Hidràulica, Marítima i Ambiental, ETSECCPB, Universitat Politècnica de Catalunya, C/ Jordi Girona, 1-3, Modul D-1, 08034, Barcelona, Spain.
E-Mail: jorge.brenner@upc.edu

Abstract

In the past decades the Catalonian coastal area (Mediterranean Sea) has been subject to an intense human pressure due to the high concentration of activities and settlements. Although some of the impacts are evident others are not, or have not been studied in detail in this region. In the present study we evaluate the environmental quality of the coastal/marine area through the coastal ichthyofauna diversity and distribution. We attempt to assess the potential influence of such activities on fish species identified as state indicators. This was done by identifying a number of land-originated indicators of pressure and relating them with their impact on fish. Special interest was paid to rare and special concern species that occur on Catalonian littoral waters.

We developed a spatial system that is based on a georelational database model and implemented on a Geographic Information System (GIS) environment. The GIS final application provides database maintenance, analysis and visualization capabilities to the information system. The data model has three areas: biodiversity, physico-chemical environment and pressure/impact descriptors. The biodiversity area is based on species presence/absence occurrences and related to their ecological and conservation attributes. The physico-chemical area is based on marine environment dynamic variables. Pressure and impact descriptors are based on different data features and produced with impact extent algorithms proposed by the authors. In the three areas, features on spatial layers were attributed with behaviours as natural as they are supposed to be found in nature and/or anthropogenic environments. Fish spatial representation and ecological data were specially related using ecological domain attributes or relations on GIS. This design provides to the database a more natural relationship among elements and environmental responses.

Keywords: Spatial database model; GIS; ichthyofauna; bioindicators; coastal zone environment.

Introduction

In its Plan of Implementation, the 2002 World Summit on Sustainable Development (WSSD) endorsed The Hague Ministerial Declaration of the Sixth Conference of the Parties to the Convention on Biological Diversity (CBD, 2003) that committed them "to achieve by 2010 a halt of biodiversity loss at global, regional and national level as a contribution to poverty alleviation and to the benefit of all life on earth" (UN, 2002). There are a number of initiatives to assess the state of the world's biodiversity who are making progress in order to accomplish the 2010 target. One of the most prominent is the Millennium Ecosystem Assessment (2004), which uses the approach of ecosystem services and is currently reviewing the status and trends of biodiversity. Critical to the success of this and similar initiatives is the development of the biodiversity informatics emergent discipline (*for a complete description see* Canhos *et al.*, 2004). This field has great potential in diverse realms, and represents the conjunction of efficient use and management of biodiversity information with new tools for its analysis and understanding.

There are several global and regional efforts that contribute to the development of biodiversity informatics by aiming at organizing data stakeholders and making data available for conservation and sustainable development research. This is the case of the Global Biodiversity Information Facility (GBIF), that since its creation in 2001 has promoted, in conjunction with its partner network (nodes), the development and adoption of standards and protocols for documenting and exchanging biodiversity data. At present, GBIF constitutes the larger biodiversity data access through its Internet portal (www.gbif.org).

Likewise, several legal and scientific motivations have emerged and lead to a better understanding of the coastal-marine environment at global and European level. The most relevant global biodiversity informatics initiative is the Census of Marine Life (www.coml.org), that constitutes "a growing network of researchers in more than 45 nations engaged in a 10-year initiative to assess and explain the diversity, distribution and abundance of life in the oceans - past, present, and future" (OBIS, 2005). In order to explain the complex ecosystem dynamics, new initiatives are focussing on the link between biological diversity and ecosystem functioning. In Europe the two main networks that deal with such activities are: the Marine Biodiversity Research in Europe (now finished; www.biomareweb.org), and the Marine Biodiversity and Ecosystem Functioning EU network of Excellence (www.marbef.org).

Geographical research on the role of marine biodiversity in ecosystem functioning normally depends on two data types related to the species in concern: taxonomic and distribution. The analysis of these attributes demand high quality standardized data, which are not always available, in contrast to data depth and breath that have become available mainly through the Internet. To our knowledge, besides scarce literature on a few species, only a few distributional data can be found on electronic resources, examples are the Ocean Biogeographic Information System (www.iobis.org), which is the information component of the Census of Marine Life project, and the Sea Around Us Project (www.seaaroundus.org) that focuses primarily on commercial species. Recently, the IUCN Red List provides comprehensive evaluation of ecology and distribution data, providing a framework for assessing the coverage of species but only a few marine species have been revised (IUCN, 2005).

As can be seen, there is a new and emerging framework development for the study of marine biodiversity and functioning of ecosystems. Our present work integrates some of the conceptual approaches of the reviewed projects in order to develop an information system spatial data model for assessing the ecological resilience of the coastal-marine ecosystems. In the long-term, this project is expected to contribute to the assessment of the ecological condition of the Catalonian and Spanish coastal zone socio-economic system.

Conceptual approach

Even though there is substantial evidence in the literature on how species diversity enhances the ecosystem functioning, most of the model development has been done in the terrestrial rather than the marine environment (*see a revision in* Loreau *et al.*, 2002). Ecosystem functioning is an abstract concept that can be conceptualized through some of its properties, being stability a relevant one and a surrogate of the system's equilibrium. Similarly, the stability of an ecosystem can be measured by its resilience. The ecological resilience measures the amount of energy required to move the ecosystem from one organized state of structure to another organized state (Holling, 1973).

Commonly, ecosystems are in a *transitional-unknown* state which ideally is moving forward to a desirable and more sustainable state. In this case, the change rate (energy needed) is a dependent function of the system's stability, which is also a function of the prevailing structure and function. These are two basic characteristics of ecosystem homeostasis across geographic scales, as presented in Fig. 1. Even though ecological resilience has not been applied frequently in ecology until now (*see work of the Resilience Alliance*: www.resalliance.org), it has been used "extensively" in its inverse form: vulnerability. Carpenter *et al.* (2001), Peterson *et al.* (1998) and other authors suggest that resilience represents in an integral manner the condition of the entire system, while structure and other functioning indicators (individually) represents only one part of it. Accordingly, if the system has more stability then it is more balanced and then it belongs to a better shape ecosystem group of the overall natural system.

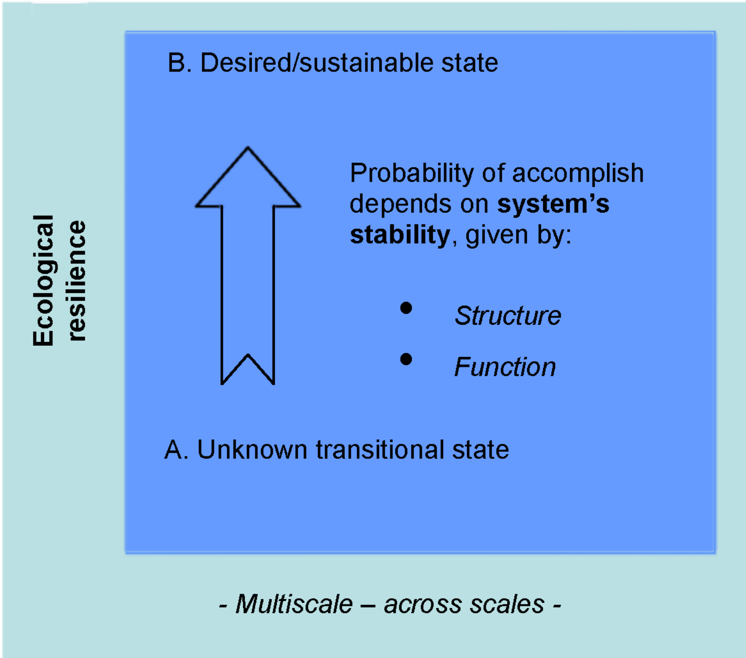


Fig. 1. Relation between ecosystem stability and ecological resilience.

Ecological resilience has three basic properties that can be used to analyse how it relates to the ecological condition of the coastal system, and therefore constructs a state indicator (adapted from Carpenter *et al.*, 2001):

- Measures the amount of change that a system can support before it moves from its stable domain.
- Proxy of the organizational? degree of the system components based in its structure and process.
- Learning process of the system to adapt to new conditions (and survive).

Ichthyofauna diversity

For the purposes described, ichthyofauna will be used as the biodiversity group to test the hypothesis. This group has been documented to provide several important functions suggested as relevant for system's resilience and thus contribute to the ecosystem's stability (Whitfield and Elliott, 2002; Holmlund and Hammer, 1999). An evaluation of the contribution of fish to the potential ecological resilience will be conducted, specifically using their diversity in *link, memory and response* functions from a macroecological perspective (ICSU, 2002; Lundberg and Moberg, 2003). *Ecological memory* is the property of fish diversity to provide essential capacity for reorganization in the form of structure and functions after a disturbance to an ecosystem, which is made possible due to the resource and process *mobile-linking* behavior of this group (Lundberg and Moberg, 2003). Complementarily, *response* is the property of the group that provides the capacity to respond differently to environmental pressures or change, as could be their reproduction diversity (Ives *et al.*, 1999; Folke *et al.*, 2002). These properties will be considered as aggregative characteristics to form functional group clusters.

The functions have been analyzed based on their role to improve ecosystem's capacity of reorganization after the disturbance. The specific structure attributes of fish to be used to form clusters or functional groups are presented in Figure 2.

LINK	• *Swimming mode* • *Max weight* • *Total length* • *Depth range* • *Environment*
MEMORY	• *Reproduction type* * • *Growth* * • *Swimming mode* • *Feeding habit* • *Total length*
RESPONSE	• *Reproduction type* * • *Growth* * • *Feeding habit* • *Depth range*

** Group of parameters*

Fig. 2. Fish functional groups and attributes.

Biodiversity data model

A data model is a basic template for implementing Information System projects, as well as GIS projects. It basically deals with data inputting, formatting, geoprocessing, sharing, creating maps, and performing analyses, among other tasks. A data model provides a general framework for writing program codes and maintaining more professional applications, and usually represents the scientific structure for data query, model, visualization and discovery (ESRI, 2004). General taxon-based data models exit widely, but little fundamental understanding of biodiversity - ecosystem functioning has been accomplished with such applications. Sustainable use of biodiversity requires a more holistic, multispecies and ecosystem-based approach, and such a level of understanding of the distribution patterns and relations among species and the ecosystem needs to be based on a robust spatial biogeographical data model (Tsontos and Kiefer, 2003).

A Biogeographic Information System mostly focuses on the integration, analysis visualization and discovery of species and environmental variables, such as biological, physical and chemical. The spatial modelling capacities of the system will be established by the data model definition and complexity, and determine its power to address one or more of the next issues:

- Biogeographic complexity
- Role of biodiversity function
- Ecosystem's ecological condition
- Conservation priorities and gaps
- Monitoring/management tool for ICZM

Biogeographic Information System

The Biogeographic Information System was developed as a georelational database model and implemented on a GIS environment. The spatial information system provides maintenance, analysis and visualization capabilities to the data bases. The system main modules are: marine biodiversity, environmental (physico-chemical) and socio-economic. Spatial features of three modules were attributed with *behaviours* as natural as they were supposed to be found in nature and/or anthropogenic environments, and structured by several geo-ecological domains as habitats, sites (conservation) and other tiling forms as distribution (species range), ecological communities and geopolitical boundaries. See module structure in Figure 3. The design characteristics provide a natural relationship among elements and environmental responses. The system also uses a metadata module compliant with the FGDC standards.

A series of vector and raster spatial features were defined as part of the catalog of the data model. Features were selected from other existing and successful implemented ESRI GIS data models as the Marine Data Model (Oregon State, Duke University and ESRI; homepage: http://dusk.geo.orst.edu/djl/arcgis/index.html) and the Biodiversity Data Model (NatureServe and ESRI) (ESRI, 2004). Spatial features used in the Biogeographic Information System are presented in Table I.

Table I. Marine GIS features used in the Biogeographic Information System (modified from Halpin, 2002).

Feature	**Fixed points**	**Instantaneous points**	**Survey points**	**Boundary lines**	**Areas**	**Time duration areas**	**Regularly interpolated surfaces**	**Irregularly interpolated surfaces**	**Volumes**
Attributes	ID X,Y Z	ID X,Y ΔZ t	ID X,Y ΔZ t1…t2	ID X1,Y1;X2,Y2… Z	ID X1,Y1;X2,Y2…X1,Y1 Z	ID X1,Y1;X2,Y2…X1,Y1 Z t1…tn	row1,col1 …rown,coln Z1…Zn	row1,col1 …rown,coln Z1…Zn	ID X1,Y1,Z1; X2,Y2,Z2 …X1,Y1,Z1 Z t1…tn
Examples	Ocurrence, conservation status	Biodiversity observation, fish densities, CTD, XBT	Samples, transects	Shoreline, EEZ, not enclosing area, legal boundaries, pressures, impacts	Habitats, distributions, patches, MPAs, pressure, impact, ecoregion, community	Oil spill, presence / absence, range	Predicted distribution, bathymetry, SST, climatology, GeoTiff	TINs, bathymetry	Front, plume, 3D habitat

The *core* biogeographic module is based on a general Species Index which serves a unique taxonomic ID for the species and functional group sub-data bases and entities. It also integrates the Element Occurrence (EO; species georeference) entities, as well as any input records from external databases. Implemented species sub-database entities are: taxonomic, conservation, ecological (observational) and information sources.

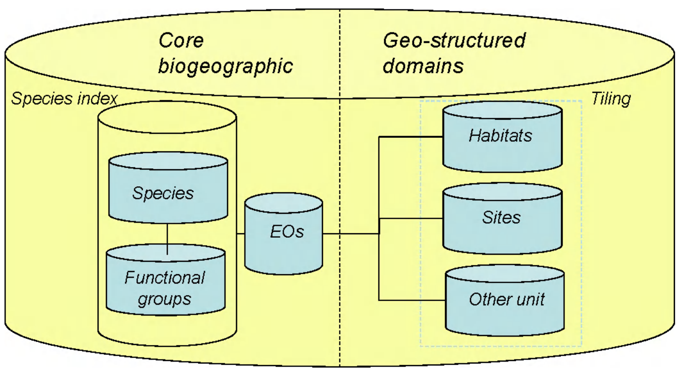


Fig. 3. Conceptual entities and relations of the biogeographic module.

Community level resilience

As a second phase *under development* of this project, we implemented the Biodiversity Information System for the assessment of the ecological condition of the Catalonian coastal zone socio-economic system. The Catalonian coast is one of the fastest developing areas in the Western Mediterranean, having more that 40% of its population (2.8 mill.) living in coastal municipalities (IDESCAT, 2004). Coastal activities, as recreation and urbanization demand large areas and are considered to be the two major sources of environmental pressures along the coast. This region is now preparing itself for the implementation of an Integrated Coastal Zone Management (ICZM) Strategy (GenCat, 2004); therefore there is an increased interest to assess the environmental services that the coastal zone provides to the global socio-economic system. This need has put scientists to work on the understanding of the properties that provide stability to the ecological systems, thus goods and services to the society.

Benthic communities and fish occurrences along the more than 800km of the Catalonian coast have been cartographied in the Biodiversity Information System. In the spatial analysis we will calculate the resilience based on the contribution of previous defined functional groups at each community. The model will also use the input of additional spatial sub-models that have been developed to study their interaction with communities' resilience to perform a *fuzzy* algorithm. See the conceptual model in Figure 4.

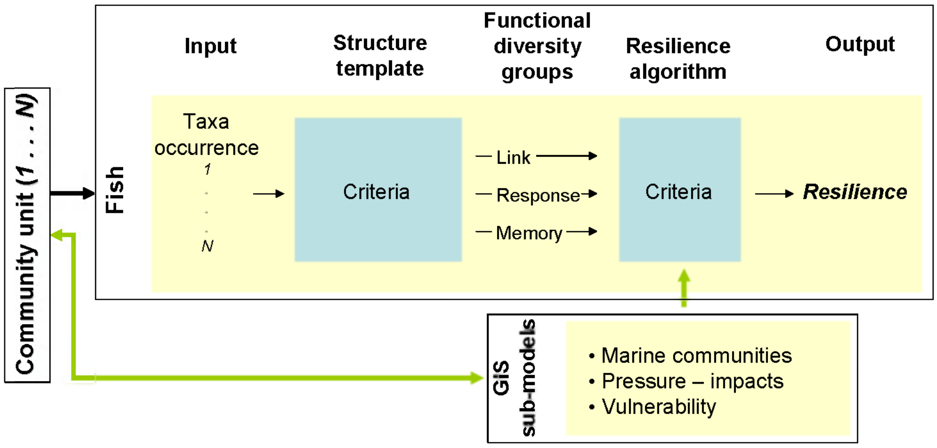


Fig. 4. Conceptual model of resilience analysis at community level.

An assessment of the total resilience of the system will be based on the individual values of the coastal benthic communities. In most areas fish is a well know biodiversity group; and thus it is expected that the resulting spatial indicators and framework could be used to promote more sustainable coastal strategies and actions at a global scale.

Conclusions

In this project we focused on biodiversity functions that contribute at system level and as biotic integrators of ecosystem dynamics, since biodiversity can provide a number of functions that has been proved to be necessary for ecosystem resilience. The resilience is enhanced by the distribution of functional groups at within scale and across multiple scales without meaning redundancy, but overlapping on time and space (complementarily).

Building resilience can be a better management strategy of *controlling* the ecosystem, in order to provide and/or maintain the ecosystem functions. A resilient ecosystem constitutes a more stable state/condition, prevents disturb-enforced change and improves the system components viability. Similarly, fish diversity is supposed to have beneficial effects on the stability of the processes of the marine ecosystem, thus the identification of taxa clusters will characterize their response to various environmental disturbances.

The developed marine biodiversity data model has integrated several designing and implementation *best practices* from existing models. Although it has only been implemented using fish diversity, we consider that a robust biogeographic data model has been developed. Furthermore, it has been designed as a platform-independent conceptual model, which provides a flexible template for other marine applications.

The development of spatial bioindicators can be better accomplished at a macro-ecological level, since high detailed occurrence or distributional data is commonly absent. We also found that the species-presence-only data in relation to the environmental factors and coastal originated human impacts is scale dependent of the biophysical model structure that has been defined. This can be integrated in a more efficient data model through the behaviour definition of the spatial features that also need to be structured by hierarchical domains. The design (modelling) of species behaviours is directly influenced by data depth, breath and quality and determines the implementation of the data conceptual model.

Acknowledgements

This paper arose from work carried out by the Coastal Management Group of the Universitat Politècnica de Catalunya and their interest on the coastal and marine environment. We are grateful to all members of the group for their useful discussions and to the Marine Engineering Laboratory for initiating and supporting the group activities. We also would like to give a special gratitude to the Fishbase Project Group for the data of Catalonian Sea fish fauna. Finally, thanks to the anonymous reviewers that helped to clarify this presentation and paper.

References

Canhos V.P., S. Souza, R. Giovanni and D.A.L. Canhos. 2004. Global biodiversity informatics: setting the scene for a "new world" of ecological modeling. Biodiversity Informatics 1:1-13.

Carpenter S., B. Walker, J.M. Anderies and N. Abel. 2001. From metaphor to measurement: resilience of what to what. Ecosystems 4: 765-781.

CBD. 2003. Handbook of the Convention on Biological Diversity. Second Edition. Secretariat of the Convention on Biological Diversity. Montréal.

ESRI. 2004. Data models. Environmental Systems Research Institute. http://support.esri.com/index.cfm?fa=downloads.dataModels.gateway. October, 2004.

Folke C., J. Colding and F. Berkes. 2002. Building resilience for adaptive capacity in social-ecological systems. In: Navigating social-ecological systems: building resilience for complexity and change. Berkes, F., J. Holding and C. Folke (Eds.). Cambridge University Press. Cambridge.

GenCat. 2004. Plan estratégico para la gestión integrada de las zonas costeras de Cataluña. Memoria Ambiental. DMAiH. Generalitat de Catalunya, Barcelona. 66p.

Halpin P. 2002. Marine GIS data model. ArcMarine: The ArcGIS Marine Data Model. Presentation downloaded from http://dusk.geo.orst.edu/djl/arcgis/index.html. October, 2004.

Holling C.S. 1973. Resilience and stability of ecological systems. Annu Rev Ecol Syst 4:1-23.

Holmlund C.M. and M. Hammer. 1999. Ecosystem services generated by fish populations. Ecological Economics 29:253-268.

IDESCAT. 2004. Anuari estadístic de Catalunya 2004. Institut d'Estadística de Catalunya. Barcelona. http://www.idescat.es/. October, 2004.

International Council for Science. 2002. Resilience and sustainable development. ICSU Series on Science for Sustainable Development. No. 3, 37 pp.

IUCN. 2005. IUCN Red List of Threatened Species. http://www.redlist.org/. January, 2005.

Ives A.R., K. Gross and J.L. Klug. 1999. Stability and variability in competitive communities. Science 286:542-544.

Loreau M., S, Naeem and P. Inehausti. 2002. Biodiversity and ecosystem functioning, synthesis and perspectives. Oxford University Press, Avon. 294 pp.

Lundberg J. and F. Moberg. 2003. Mobile link organisms and ecosystem functioning: implications for ecosystem resilience and management. Ecosystems 6:87-98.

Millennium Ecosystem Assessment. 2004. Millennium Ecosystem Assessment. http://www.millenniumassessment.org. November, 2004.

OBIS. 2005. Ocean Biogeographic Information System. http://www.iobis.org/. January, 2005.

Peterson G., C.R. Allen and C.S. Holling. 1998. Ecological resilience, biodiversity, and scale. Ecosystems 1:6-18.

Tsontos V.M. and D.A. Kiefer. 2003. The Gulf of Maine biogeographical information system project: developing of a spatial data management framework in support of OBIS. Oceanologica Acta 25:199-206.

United Nations. 2002. Report of the World Summit on Sustainable Development. Johannesburg, South Africa.

Whitfield A.K. and M. Elliott. 2002. Fishes as indicators of environmental and ecological changes within estuaries: a review of progress and some suggestions for the future. Journal of Fish Biology 61 (Supp. A):229-250.

MEDAS system for archiving, visualisation and validation of oceanographic data

V. Dadic and D. Ivankovic

Institute of oceanography and fisheries, Split Croatia, Mestrovicevo setliste 63, 21000 Split, Croatia
E-mail: dadic@izor.hr

Abstract

The Marine Environmental Database of the Adriatic Sea (MEDAS) serves for data management of different types of oceanographic parameters. The system was constructed to capture historical as well as real-time data and to make them available through the World Wide Web. A special subsystem was developed for validation of various oceanographic parameters using IOC protocols (IOC, 1993), as well as transforming data from/to different international formats.

Keywords: oceanographic database; historical and real-time data; data validation; data assimilation; numerical models.

Introduction

In the framework of the Croatian National monitoring programme of the Adriatic Sea MEDAS was developed using Oracle RDBMS, Java and ArcGIS tools. Its core is a relational database that includes general information on oceanographic data, responsible institutions, persons, etc. Further thematic subdatabases with measurement data related to physical and chemical oceanography, biology and fisheries also form part of the system (fig. 1). MEDAS has improved our management of oceanographic data, including archiving, dissemination, validation, visualisation and presentation of data through web pages in both text and graphical formats (Dadic *et al.*, 1995; Ivankovic, *et al.*, 2000).

A large volume of diverse oceanographic data has been stored in the database (more than a million records). As they have been measured during more than a century, their distribution in space (geographic position and layer in the water column) and time (year, season and month) is haphazard (Zore-Armanda *et al.*, 1991; Levitus and Boyer, 1998; MEDAR-MEDATLAS, 2002). Therefore, it was necessary to develop a procedure for data quality control and appropriate validation flagging prior to data processing.

Retrieving data from MEDAS is simple, so the user can retrieve any data in the following two steps:

- basic information about measured parameters (what, where, when, who, etc) can be obtained by searching the relational database

- through a connection to the thematic databases to get information through various menus. Depending on permissions, the user can access, insert or update records.

Creation of the database

The development of databases related to marine research is a specific task, in the first place because of the special requirements of oceanographic science. Users need to have easy and straight-forward access to the data, their derived parameters and their practical application. The basic considerations for the creation of marine environmental databases and secondary programmes are as follows:

- The majority of data have spatial and temporal components. The temporal component is simple to query (query period from-to), but for the spatial component it is necessary to develop a map interface to select data for analysis of a particular area
- The possibilities of research in a particular area depend on the quality of the data. To reach high-quality data, both human and material resources are needed. Data must be well protected from corruption and misuse, but without hampering ease of access for users
- While collecting data, it is possible to make different mistakes at various points. Hence, it is very important to develop constantly improved qualitative validation of data that will point to possible mistakes, hereby leaving the possibility for a final decision by the user
- To report on collected data, different processes and procedures have been used. It is very important to store these constantly changing procedures into a database. That way faster reporting can be achieved, and more than one model can be applied at the same time
- To present the results from the analyses it is necessary to develop a graphical presentation in order to better show the interdependence of different factors considered. The results have to be presented in their spatial context
- The database interface must be simple, accessible and, if is possible, platform-independent. Connection trough the Internet is a big advantage, and at present becomes imperative.

For all the above-mentioned points, it is obvious that it is a complex task that requires a multidisciplinary approach. It is often a never-ending process, and it demands constant step-by-step improvement to increase the usability of the database. It requires frequent feed-back from the end-user: there is the possibility of not having satisfactory results due to a too wide and theoretical approach.

Tools and technologies

The MEDAS database and web interface are based on Oracle 9i RDBMS and Application server. Java applets are used as mapping tools and for data visualisation. Special applications for automatic inserting and processing of real time data were developed in the C++ program language. Oracle Graph and Matlab are used for data visualisation. The rationale behind these choices was as follows:

- Oracle 9i: Stability, security; easy installation and migration
- Oracle Application Server: no need for client side software (only browser), accessibility (different institutions, ship – GPRS), easy web publishing (default)
- Java applet – mapping tool and data visualisation: portability (cross platform), use of client-side resources (no need for powerful server)

These tools also have some disadvantages, such as software license cost, instability and bugs of some Java versions and sometimes need for manual JVM installation. In addition, the interface for transcoding data to ArcGIS 8.1 input for a spatial presentation of data was developed.

Database design

Database design is a critical stage in the whole process of database development since the performance of the database depends upon its design. All required database objects should be carefully planned and co-ordinated at the beginning of the development. All the previous experience from that field is very welcome.

As it can be seen in figure 1, the main entity is 'measurements': the chart contains the facts about time and depth, while the other facts are gained from other entities. The measurement results are aligned according to smaller entities, into sub-tables that are referred to the measurement table. Figure 1 shows the entities so far. Further expansion of the number of entities (tables) with concrete data for other types of measurements (heavy metals, biochemistry, etc.), is planned.

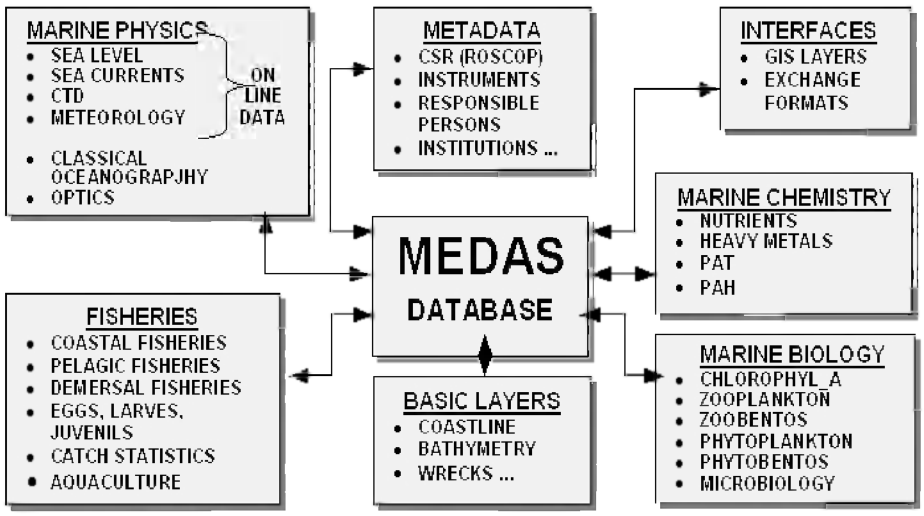


Fig. 1. The most important objects in database.

Database capabilities

The MEDAS system is used in the framework of the Croatian national monitoring programme (Systematic Research of the Adriatic Sea as a Base for Sustainable Development of the Republic of Croatia), in which various Croatian research institutions are involved. The database is used for storage, validation, and presentation of real-time measurements, and for running circulation models. In addition, the database is used for near real-time processing of data from research vessels. Visualisation and mapping tools are important parts of the database for both quality control and web presentation of the data. Various types of import and export formats are very useful for data sharing and processing. Figure 2 shows the main groups of MEDAS functionalities.

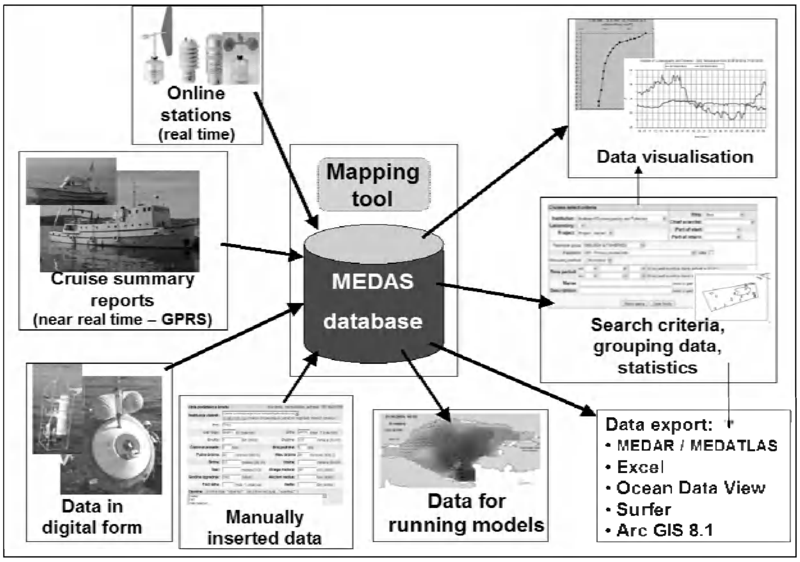


Fig. 2. Database functionalities.

CSR (Cruise summary report)

The CSR web interface includes forms for data input and also a java applet as mapping tool. The objective is to develop a fully web-based database interface. The advantage of a web-based database interface is that is platform-independent, and only requires a browser and access to the Internet. With this facility we can insert CSR data directly from research vessels (GPRS, satellite internet connection) and from various different institutions. The system contains procedures for automatic checks of CSR data (time-spatial crosscheck). All CSR data are available online (http://www.izor.hr/roscop/eng/) with detailed information about time, station position, measured parameters, responsible persons, projects, institution and ship details. Cruise information includes a mapping tool that is implemented using frames with a form – applet interaction. CSR data are linked with measurements; together with an interface for authorisation, this system is used for data quality control and processing.

Real-time data

The real-time part of MEDAS includes programmes for automatic upload of data from real-time meteo-ocean stations, and an automatic visualization and Internet publication. On a daily basis, model results are automatically calculated and published on the Internet. The real-time part is available online at http://www.izor.hr/eng/online/. Figures 4 to 7 shows some web examples of real-time and aggregated data.

Real-time data are used as input for circulation models. For example, the model assimilates data from Punta Jurana online station. The model is initiated with hydrostatic conditions; after that, the results from any day serve as initial conditions for the following day. Visualization of model results is generated by a C++ application. The first application generates a dynamic Matlab script, which is run, to dynamically create the html pages. Some examples are shown on figures 8 to 10.

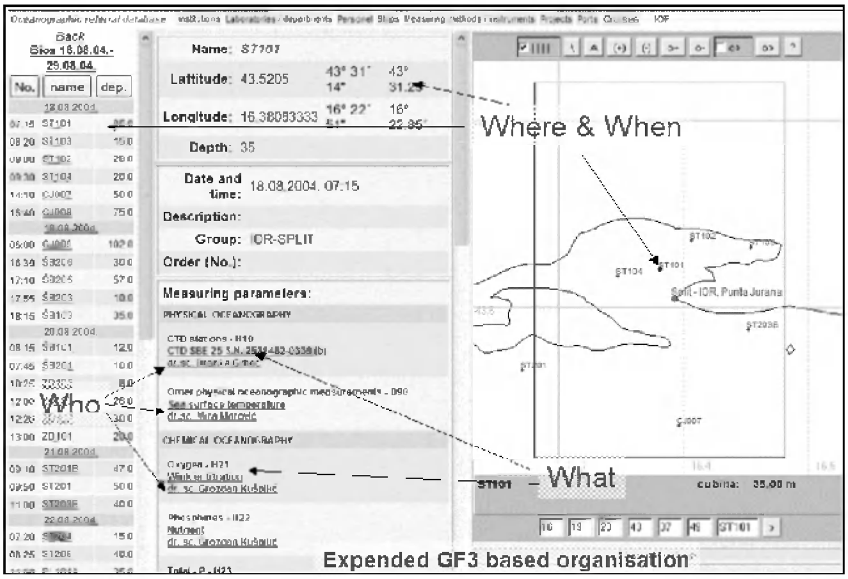


Fig. 3. Example of extended on-line CSR (browser).

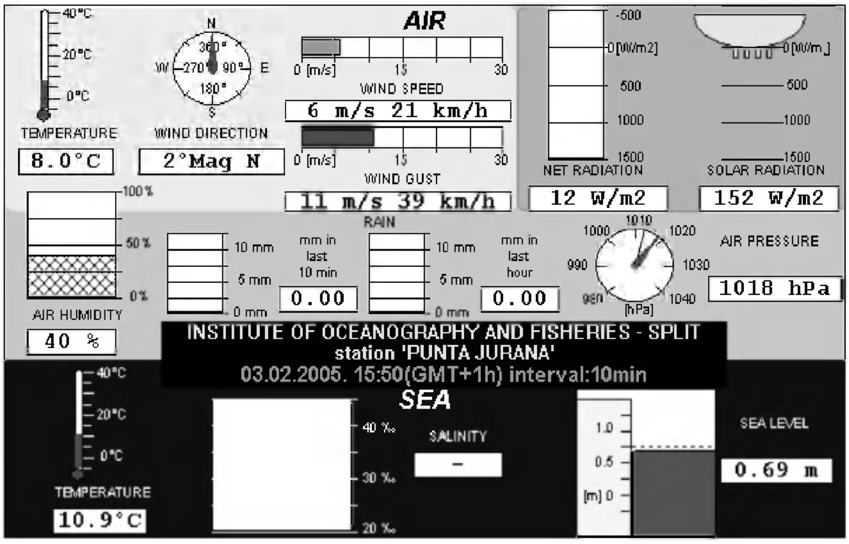


Fig. 4 Visualisation of real-time data.

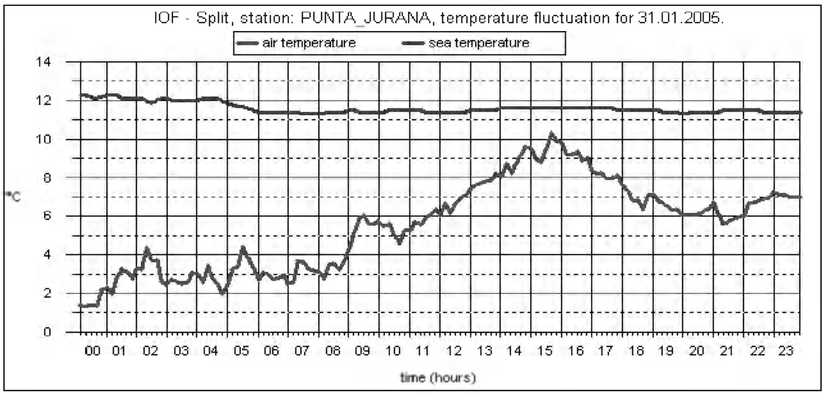


Fig. 5. Graphical presentation of 24h air pressure and humidity fluctuation.

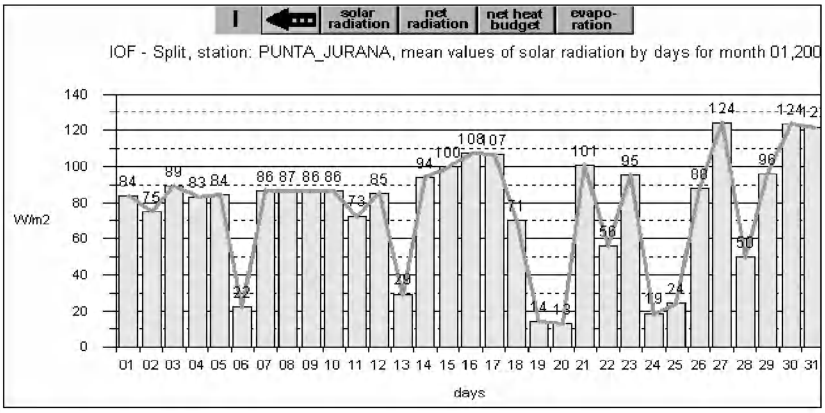


Fig. 6. Monthly mean solar radiation per day.

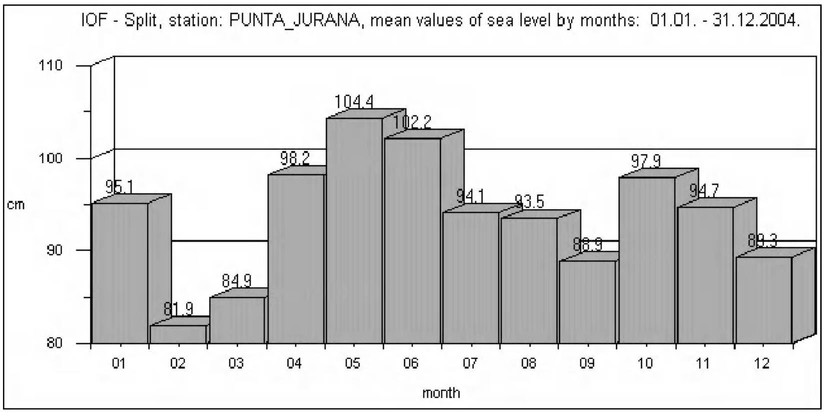


Fig. 7. Yearly mean air temperatures per month.

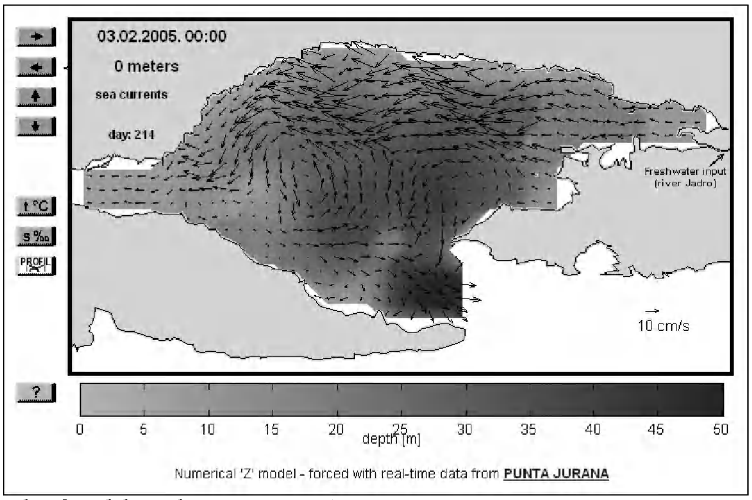


Fig. 8. Example of model results (sea currents).

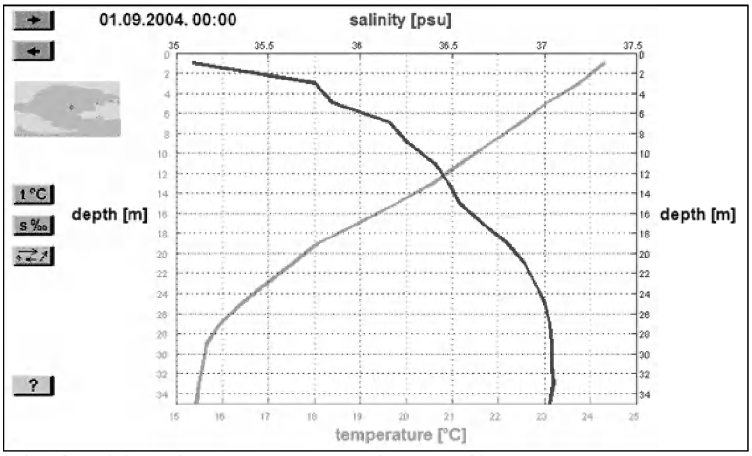


Fig. 9. Example of model results (temperature – salinity profile).

Tools for data validation

Oceanographic data stored in MEDAS are randomly distributed in space (geographic position and layer in the water column) and time (year, season and month). Therefore, it was necessary to calculate monthly and seasonally climatological values of each parameter (mean value, minimum, maximum and standard deviation) at standard oceanographic depths and specific marine areas, which most often are defined as squares. As there is a large amount of data in our database, manual validation is practically impossible. Therefore, based on protocols of IOC-UNESCO (IOC, 1993) and World Data Centre (Boyer and Levitus, 1994) a procedure was defined, which includes comparison of different parameters and an iterative method following several steps:

- Calculation of monthly and seasonal means and standard deviations from all data for each parameter at the same measuring level for the entire marine area
- Comparison of each reading with the calculated mean value and its exclusion from processing if the difference exceeds a predefined range (from 1 to 5 standard deviations, depending parameter and entire area)
- Interpolation of values from observed levels to standard oceanographic levels
- Comparison of each interpolated value with the calculated mean value and its exclusion from processing if the difference exceeds a predefined range (from 1 to 3 standard deviations depending parameter and entire area)
- Visual check of profiles based on interpolated values at standard levels
- Estimating interpolated data in square nodes of geographic grid
- Presentation of output results in graphic form and check for 'bull's eyes'
- Repeating procedure if necessary

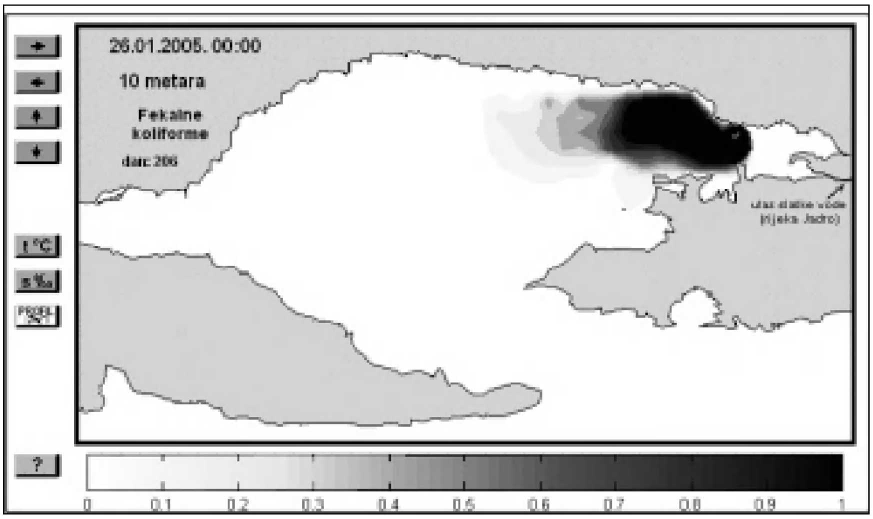


Fig. 10. Reconstruction of faecal coliform field in Kastela Bay obtained by Z-model with assimilation of various data from the station Punta Jurana.

Interpolation of values on standard oceanographic levels was done by third order Newton method of finite differences modified by referral curves (Reinger and Ross, 1968, Dadic *et al.*, 2002). This method is specially recommended for oceanographic parameters sampled at the discrete oceanographic levels (*e.g.* by water samplers and nets). Ordinary kriging method and semi-variogram were used for spatial analysis of data (Journel and Huijbregts, 1995).

Climatologic values derived by above procedure serves for validation of new data received in the database. This process cannot be fully automated and expert opinion is needed to decide on the quality of each datum after checking by visualisation. Depending on quality, each datum is assigned a quality flag (MEDAR-MEDATLAS Group, 2002).

As the main objective of any spatial investigation is to simplify the analysis of spatially-distributed data to end users, a special interface between Oracle and ArcGIS 8.1 tool was developed. This system enables overlaying layers of different parameters and their presentation through maps. Based on these maps data validation and various analyses of oceanographic properties of the Adriatic Sea have been possible.

Some results of data analysis

As a result of data analysis using the MEDAS system, many duplications, uncertainties and erroneous historical data were identified. For example, about 49.7% of BOT data were duplicated, 3.2% outside climatologically range, 0.7% of oceanographic stations were attributed to wrong position, and about 17.4% of BOT data attributed as MBT data.

Based on the data analysis, four different sub-regions of the Adriatic Sea with similar oceanographic properties were recognised and 41 standard oceanographic levels defined as suitable for climatological analysis.

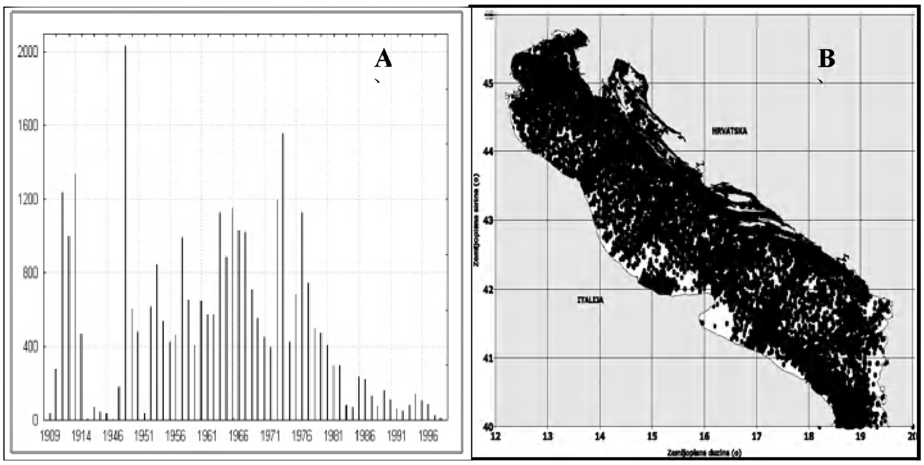


Fig. 11. Number of measured temperature profiles by year (A) and spatial distribution of station in the Adriatic Sea (B).

After validation and harmonisation of data, numerous unique and corrected data profiles were included in statistical calculations, *e.g.* more than 100,000 temperature readings. The importance of validation procedures and harmonisation of classical oceanographic observations is shown in figure 12. The temperature spatial field gathered from original data looks very artificial, but looks much more realistic when calculated from partly validated data. Many of the unrealistic features, like very high gradients and bull's eyes, disappeared from the objective analysis after the data validation.

Based on climatology analyses of more oceanographic parameters (temperature, salinity, oxygen, nutrients, chlorophyll_a), 41 standard levels and two maximum depth difference criteria (inner and outer) were defined for each standard level (Dadic *et al.*, 2002).

As there was anisotropy in spatially distributed data in transversal and longitudinal axes of the Adriatic Sea, different radius influence and grid mesh were used. So, influence radius of 7.5km and interpolated data on a 5 x 5km grid were used in transversal direction of the Adriatic Sea and in the coastal area and influence radius of 12km and 7.5 x 7.5km grid in longitudinal direction at the open sea during kriging interpolations.

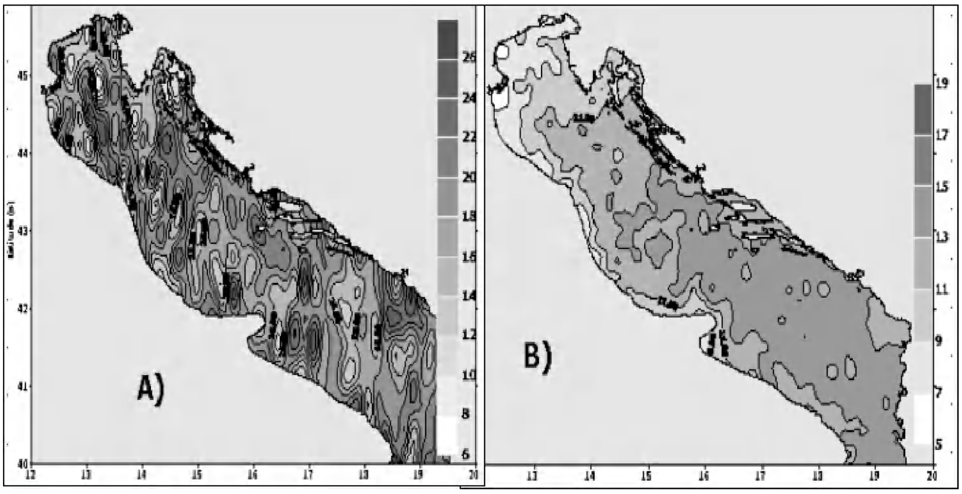


Fig. 12. Interpolated field of mean value of temperature at the sea surface gathered from the original data (A) and data passed quality control (B).

References

Boyer T. and S. Levitus. 1994. QC and processing of historical oceanographic temperature, salinity and oxygen data. SilverSpring. USA-NOAA technical report NESDIS, No. 81, 112p.

Dadic V., M. Srdelic and A. Grubisic. 1995. Oceanographic information system and databases of the Adriatic Sea. *Proc. of 17th International Conference ITI'95,* Pula, 289-294.

Dadic V., M. Srdelic and Z. Grzetic. 2002. Analysis of oceanographic properties of the Adriatic Sea by GIS technique. Cartography and information 1:46-59.

IOC. 1993. Manual of Quality Control Procedures for validation of Oceanographic Data, prepared by CEC-DG: XII, MAST and IODE. Paris. IOC (UNESCO) Manuals and Guides NO. 26:436p.

Ivankovic D., V. Dadic and M. Srdelic. 2000. Marine environmental database of the Adriatic Sea with application for managing and visualisation of data. ENVIROSOFT International Conference: Computer techniques to Environmental Studies VIII, 28-30 June 2000, Bilbao, Spain. Proceedings:398-406.

Journel A.G. and Ch.J. Huijbregts. 1993. Mining geostatistics, Academic Press, New York-London. 600pp

Levitus S., and T. P. Boyer, 1998. World ocean atlas 1988. SilverSpring. NOAA Atlas NESDIS 4, 117 p + 7 CD.

MEDAR-MEDATLAS Group. 2002. Mediterranean and Black Sea database of temperature, salinity and bio-chemical parameters climatological atlas. EU-MAST Programme. Brest. IFREMER Ed. 4 CD.

Reinger R.F. and C.K. Ross. 1968. A method of interpolation with application to oceanographic data. Deep-Sea Research 15:185-193.

Zore-Armanda M., M. Bone, V. Dadic, M. Morovic, D. Ratkovic, L. Stojanovski and I. Vukadin. 1991. Hydrographic properties of the Adriatic Sea in the period from 1971 through 1983. Acta Adriat. 32 (1):1-547.

Constraints, Suggested Solutions and an Outlook towards a New Digital Culture for the Oceans and beyond: Experiences from five predictive GIS Models that contribute to Global Management, Conservation and Study of Marine Wildlife and Habitat

Falk Huettmann

419 Irving I, EWHALE lab, Institute of Arctic Biology, Biology and Wildlife Department,
University of Alaska-Fairbanks, Fairbanks AK 99775 USA
E-mail: fffh@uaf.edu

Abstract

Marine wildlife and habitat data are increasingly available to the global public for free and via the internet. This 'data explosion' brings change in ocean management and promotes predictive modelling. Predictive modelling using such data and GIS (Geographic Information Systems) has matured as a robust research method, but still does not get used to its full potential. This study reviews experiences and constraints encountered during 5 predictive GIS models representative for the Atlantic and Pacific. It was found that data availability is less of a problem, but data quality still needs to be improved in time and space. Bigger constraints were found with the management and policy implementations of spatial models. A professional attitude towards the free delivery and use of data, and data availability is required. Often, expertise and skill is still missing on how to set up, build, interpret and implement predictive models towards safeguarding marine wildlife and its habitat. It is suggested that the awareness, education and support for data and modelling needs to be further improved in the public, in agencies and among scientists. Using evaluated models should become a legal requirement when dealing with endangered wildlife and habitat of the global village. A change towards a truly digital and transparent administration and culture, based on science-based management and using models for decision-making, is suggested for the oceans and beyond.

Keywords: Geographic Information Systems (GIS, predictive modelling, databases, marine wildlife and habitat).

Introduction

The rapid increase in data availability for the oceans brings changes. For instance, it affects decision-making and supports spatial and predictive modelling of wildlife species and their habitat. Predictive modelling is a relatively new but already mature research discipline which is still on the rise. Modelling high quality data contributes to conservation, management, research and to a sound decision-making in a complex and fast changing world (Ford, 1999; Sarewitz *et al.*, 2000; Shenk and Franklin, 2001). Often, predictive modelling represents the only method to obtain sound information for marine wildlife and its habitats in larger areas, *e.g.* when only opportunistic samples exist in space and time. This is specifically the case when study areas are large, remote or difficult to access such as coastal areas, pelagic habitats and oceans. However, when trying to apply these predictive modelling methods in the real world and in policy it becomes quickly obvious that major constraints beyond the technical possibilities still exist. From earlier applications elsewhere it was shown that data availability has been the major constraint (Huettmann, 2000a, 2004; Esanu and Uhlir, 2004; Gottschalk *et al.*, 2005), but many examples nowadays can be found where the ocean has received great data projects representing a progressive template for other ecosystems regarding data availability, *e.g.* World Ocean Atlas (WOA, Levitus, 1994), Reynolds fields (http://podaac.jpl.nasa.gov/cgi-bin/dcatalog/fam_summary.pl?sst+), Ocean Biodiversity Information System (OBIS, http://www.iobis.org/; Malakoff, 2003; Zhang and Grassle, 2003). More relevant constraints are still brought by political influences or by traditionally trained field workers, researchers, managers and other groups either not familiar with spatial models and their interpretations or having vested interests (Huettmann, 2005). Many predictive wildlife modelling references exist, either dealing with how to perform statistically accurate modelling (*e.g.* Manly *et al.*, 2002), how to link them with biological mechanisms (*e.g.* Nakazawa *et al.*, 2004), or apply them in terrestrial applications (*e.g.* Scott *et al.*, 2002), but less so with ocean-wide, large-scale marine wildlife and biodiversity (*e.g.* Valavanis, 2002; but see Rozwadowski, 2002). Wildlife and habitat modelling techniques are complex and require multidisciplinary approaches; they often have to consider many aspects of humans and human behaviour as well in order to be successful (Huettmann, 2004).

In order to complement and further improve the existing and traditional information about marine wildlife with advanced modelling, here I present and analyze some experiences from representative modelling and model building projects in the Atlantic and Pacific dealing with a variety of marine conservation topics and marine wildlife species. Specifically, I outline issues which still need to be overcome towards more progressive and science-based management modelling in order to safeguard the natural wildlife and habitat resources of the global oceans (*e.g.* in an adaptive management framework; Walters, 1986). The presented model projects are using free data and are based on progressive and multidisciplinary studies. All of which have a field work component and where modelling contributes new insights and guidance for science and for the management process. Most of the studies discussed here try to model species habitat relationships and to predict spatial distributions, populations and future habitat states. However, for completeness, issues such as population modelling and other topics related to marine modelling also get addressed.

Methods

In the following, I describe model data sets and individual methodologies from five selected modelling projects which can get considered as a representative set of predictive ocean species models. This allows drawing general conclusions for improving modelling exercises world-wide. All of the data mentioned here refer to GIS-layers in Arc View 3.3.

Case study 1: Pelagic Seabird Species and Colony Distribution in the Northwest Atlantic.

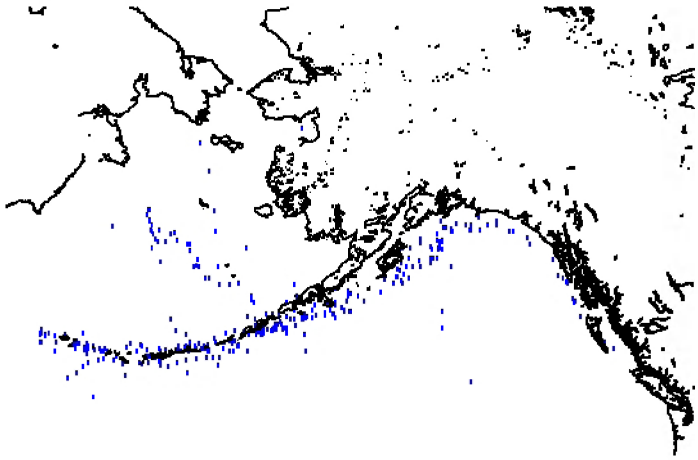
Seabird distribution of four abundant species (Northern Fulmar *Fulmarus glacialis*, Atlantic Puffin *Fratercula arctica*, Common Murre *Uria aalge* and Northern Gannet *Morus bassanus*), derived from pelagic surveys carried out during more than 25 years (1966-1992) in the Northwest Atlantic (Gulf of Maine – Canadian High Arctic) were related to marine features. Marine habitat data were available for free *e.g.* from the internet for Comprehensive Ocean-Atmosphere Data Set (COADS http://www.cdc.noaa.gov/coads/), World Ocean Atlas (Levitus, 1994), ETOPO5 and others. Seabird survey data were provided in a digital format by T. Lock and R.G.B. Brown, Canadian Wildlife Service (for more details on data and methods see Huettmann and Lock, 1996; Huettmann, 2000a). These seabird-habitat relationships were quantified using primarily a multiple regression approach predicted to locations with a known set-up of environmental features, and which get finally evaluated for its performance. More details can be found in Huettmann and Diamond (2001).

*Case study 2: Marbled Murrelets (*Brachyramphus marmoratus*) in coastal Old-Growth Forest habitat of British Columbia, Canada.*

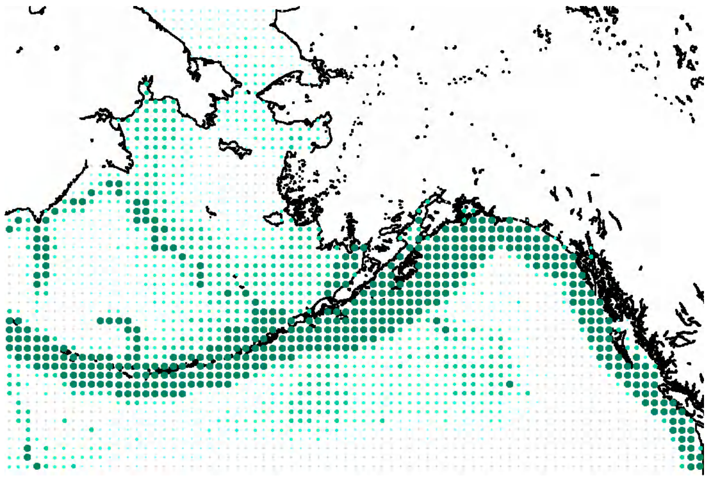
Marine abundance and potential nest occurrence information of the Marbled Murrelet, a seabird species of international conservation concern, were related to the marine and terrestrial features. Marbled Murrelet data came from Burger (1995) and other published sources; environmental data were taken from NOAA (National Oceanic and Atmospheric Administration Coastwatch, 2000) and others. The habitat preference was quantified using linear and non-linear models. These statistical relationships were then predicted to coastline locations with a known set-up of environmental features. Secondly, population estimates were also derived from these spatial models. More details about this study can be found in Yen *et al.* (2004).

Case study 3: Predicting the pelagic distribution of Short-tailed Albatross (Phoebastria albatrus) in the Northern Pacific using 'Presence Only' data.

Compiled opportunistic and historical albatross sightings ('presence only' data) already provided for free on the internet (http://www.iphc.washington.edu/staff/tracee/shorttail.htm; see Figure 1a) were used to describe the ecological niche of an endangered seabird, the Short-tailed Albatross. The primary focus of this study was to describe from Alaskan sightings as training data the distributional range of this species in the adjacent Russian and Canadian waters and for which only very few or none sightings and incomplete information exist (Figure 1b; FH unpublished).



a) Dots indicate a sighting; usually during July and August from 1940-2000.



b) Size and intensity of point indicates magnitude of the index of occurrence across months and years.

Fig. 1. a) Raw sightings ('presence only'), and (b) predictions of Short-tailed Albatross distribution throughout the year in the Northern Pacific using MARS algorithm.

Case study 4: Modelling the distribution of nesting waterbirds in the subarctic Great Slave Lake using opportunistic sightings.

The Great Slave Lake presents a major and large waterbody in the North American subarctic. Due to its general inaccessibility, it is widely unsurveyed and only limited information on nesting waterbirds exist, published in the grey literature. Opportunistic surveys (Sirois *et al.*, 1995) were used trying to derive nesting/colony distribution and abundance estimates of nesting pairs. Environmental data were used from topographical maps, climate models and remote sensing imagery. The goal of this study was to improve distribution information, to assess scale effects and to obtain first population indices for this otherwise widely unsurveyed area. Further details regarding this study are found in Fenske (2003).

Case study 5: Modelling the future coastal ecosystem of Marbled Murrelets to assess spatially explicit impacts on distribution and abundance in British Columbia.

This study is currently 'in progress' (FH *et al.* unpublished) and deals with a major contribution brought by predictive modelling: Forecasting the state of habitats for an endangered species. It is based on the initial model study presented by Yen *et al.* (2004), and tries to replace the current habitat layers - marine and terrestrial - with future scenarios in order to forecast eventually a distribution and population estimate for Marbled Murrelets for entire British Columbia and beyond. The ‘future’ is defined as 10, 50 and 100 years from present (see also Huettmann *et al.*, 2005 for methods). This project will allow obtaining a spatial Population Viability Analysis (sPVA) for a seabird species that became of international conservation due to the ongoing habitat degradation, *e.g.* logging of the old-growth forest nesting habitats, and disturbances in the marine environment.

Results

The following section summarizes the key components and experiences from each of the five models.

Case Study 1

<u>Major Contribution of the Model:</u> Results from this model present for the first time a consistent seabird distribution map which covers the entire North West Atlantic (compare with Brown *et al.* 1975, Brown 1986 for shipboard observations).
<u>Modelling Method:</u> Generalized Linear Models (GLM) and Classification and Regression Tress (Cart-SPLUS) were used.
<u>Constraints encountered during the Model Building Process:</u> Data quality of seabird and environmental data was coarse. Detailed knowledge about the biologically available habitat for seabirds was missing. Earlier views from experts of ‘how seabirds would respond to the marine environment’ biased the model building and model testing initially, and had to be overcome and revised. Lack of an interdisciplinary research

environment and software technology, including technical and administrative infrastructure problems, presented delays to complete the project efficiently.
Constraints due to Data Accessibility: None, since internet was used (for environmental data); seabird data were efficiently provided by governmental agency. Some seabird data had to be added, and restored from back-ups and hard copies; quality checks were required.
Constraints encountered during Model Implementation: The governmental management process did not consider results from modelling for more than five years. Local expertise is missing to comprehend and implement findings from these models.

Case study 2

Major Contribution of the Model: The resulting distribution map allowed for the first time for a consistent distribution information of Marbled Murrelets for the entire coastline of British Columbia. These estimates were derived from consistent methods and data, and also allowed for the first time for a modelled population estimate, obtained from compiled, best scientific available information for this species of major conservation concern.
Modelling Method: GLM, Cart-SPLUS, CART-Salford, Multiple Regression Splines (MARS-Salford) and Neural Networks (SPLUS) were applied.
Constraints encountered during the Model Building Process: The data quality of Marbled Murrelet abundance and locations, as well as the environmental data was coarse; Metadata of the Marbled Murrelet data did not exist. Knowledge about available habitat for seabirds was missing. Initial views by experts of 'how Marbled Murrelets would respond to the marine and terrestrial environment' and at what scale biased model building severely and had to be overcome (Huettmann *et al.*, in review). Political views about Marbled Murrelet research complicated and delayed the project and data availabilities strongly.
Constraints due to Data Accessibility: A centralized database of all known Marbled Murrelet nests and abundances was not available (but see http://www.sfu.ca/biology/ wildberg/ species/ mamu.html); many data sets had to be located, assessed, digitized and merged from numerous individual contractors and data holders who work small scale but lack seeing the large picture. Alternative data sets had to be obtained from NGO and internet sources.
Constraints encountered during the Model Implementation: Management process did not consider results from modelling, yet. Counter models were initiated and used to circumnavigate findings from this model. The use of AIC (Burnham & Anderson 2002) for model selection, instead of traditional significances and p-values, created a major problem in the acceptance of results from this model. Local expertise is missing to comprehend and implement model findings.

Case study 3

Modelling Contribution: For the first time, a pelagic distribution map of Short-tailed Albatross in the Northern Pacific was predicted.
Modelling Method: MARS-Salford was applied to 'presence only' data.

Constraints encountered during the Model Building Process: Local experts and governmental agencies claimed a monopoly for dealing with this species and discouraged large-scale model-building to this very day. Due to competitive funding and internationally pending legal conservation tensions the modeller was threatened and marginalized for going ahead building predictive models on Short-tailed Albatross for international peer-reviewed research publications. Lack of funding to build model, compile and work up data had to be overcome.
Constraints due to Data Accessibility: None; all data are freely and fully available on the internet/WWW.
Constraints encountered during the Model Implementation: So far, the model was ignored by the Short-tailed Albatross research community, as well as by governmental and other agencies with a mandate to manage seabirds.

Case study 4

Modelling Contribution: For the first time, a consistent distribution and abundance information of waterbirds in the subarctic Great Slave Lake was produced.
Modelling Method: GLM, CART-Salford, MARS-Salford and Neural Network SPlus using 'Presence Only' data.
Constraints encountered during the Model Building Process: Governmental agency did not fully collaborate; otherwise, no relevant constraints were encountered.
Constraints due to Data Accessibility: Data were not available or known for this project and had to be compiled, created and digitized.
Constraints encountered during the Model Implementation: The model was not considered by governmental agencies for conservation and management actions, yet.

Case study 5

Modelling Contribution: Future Marbled Murrelet habitat, distribution and abundance.
Constraints encountered during the Model Building Process: Some landowners did not provide growth and yield information, nor were they to motivate buying into an overall and mutually accepted modelling approach. Lack of data accuracy was used to block and delay the modelling process. Missing funding and seeing the importance of this work by governmental agencies had to be overcome.
Constraints due to Data Accessibility: Due to the lack of 'buy-in', landcover data as the crucial source for model building were constantly criticized.
Constraints encountered during Model Implementation: Competing models were developed from opposing lobbies on a smaller scale, presenting their own models and views into the political discussion.

Discussion

The review of modelling studies for marine wildlife and habitat shows that some consistent constraints occur within predictive modelling projects, harming crucial progress on this subject. These constraints have not been shown or explained and outlined in earlier modelling publications. Considering that modelling and its importance will increase I believe it is very important to outline and review modelling constraints for a wider audience in order to address them. Earlier topics important enough to halt entire modelling projects such as GIS, software analysis code and data availability were not considered as a major constraint anymore (see also Huettmann and Linke, 2003, Esanu and Uhlir, 2004, Gottschalk *et al.*, 2005). Instead, subjects related to the lack of technical and statistical expertise by implementing agencies, biased expert views or vested interests were mentioned most often as constraining modelling projects and their acceptance (see also Rozwadowski, 2002 for policy applications and political influences on science-based models). Topics like data quality (*e.g.* content and spatial) and data transfer/copyrights were mentioned less often, but still could block modelling work dramatically for charismatic and important wildlife species and biodiversity in general (see Graham *et al.*, 2004 for terrestrial biodiversity applications); it impairs the general acceptance of models. Spatial predictive modelling is often the only means to provide estimates of marine wildlife distribution and abundances, *e.g.* in pelagic and coastal wilderness areas that are difficult to access (Huettmann, 2000b). The advantage of modelling is that it is derived from a consistent and transparent methodology, that it can be repeated (=evaluated by other parties), and its performance assessed (Fielding and Bell, 1998; Ydenberg, 1998; Pearce and Ferrier, 2000) towards a better scientific understanding and higher trustworthiness in the management process and for public policy. I feel that these steps provide a major argument in favour of building and applying models. Once a model has been build, and a modelling culture is set up, poor models can always, and relatively easy, be improved, *e.g.* in the framework of scientific hypothesis testing. Considering such a situation and the major contributions that can be obtained through the use of predictive modelling, it is surprising to learn that the use of spatial modelling in conservation management is still not well advanced and not used more effectively (Bookhout, 1994; Primack, 1998, but see Walters, 1986; Brown *et al.*, 2000 and MARXAN http://www.ecology.uq. edu.au/index.html?page=20882 for Marine Protected Areas MPAs), nor is it built in as a requirement into the legislation of endangered species and habitat (see for instance Czech and Krausman, 2001) or in the Ocean Act and organizations administrating oceans of the world (Rozwadowski, 2002 for ICES).

Despite the experiences from the models presented here, one might find of interest as well the numerous modelling projects which eventually could not be carried out due to various constraints. At least six of such modelling projects come to mind to the author; they usually failed due to data access issues from individuals with an interest in the data themselves. Other constraints were caused by the general lack of support, *e.g.* financial and man-power, for collecting and digitizing data, for building models and for evaluating them statistically. Although financial constraints exist, other reasons for failing predictive modelling projects are brought by poor data quality and lack of awareness on the benefits of modelling, *e.g.* beyond borders. Besides failed projects, one should also consider the tremendous delay of model projects caused when data and model issues occur. One problem is for instance that even within governmental agencies, data are sometimes not well known, documented with Metadata, heavily delayed, not shared or plainly not available. Vested interests brought by promotion/salary, money/fieldwork funds and publication rights further proof counterproductive to modelling and its exciting advances for the global village.

All models reviewed, as well as many others in the literature (Manly *et al.*, 2002; Scott *et al.*, 2002), primarily deal with correlations but less with the true biological mechanisms to explain wildlife distribution and abundance. This is less of a technical modelling issue but more a data issue since biologically meaningful marine biodiversity and prey information, *e.g.* benthos, plankton and non-commercial fish databases collected with a consistent protocol are often still missing. The modellers should be more explicit in requesting these crucial data sets in order to further improve biological models and predictions.

The author found that advances of scientific exploration and innovation, such as represented by modelling, can be constrained by administrative hierarchy, and most importantly, by old-fashioned peer-review policies of grants and publications. Old-fashioned hard copy project reports do not prove helpful, but the underlying digital data are needed as well. Also, funding agencies are able to reduce global progress severely for advanced problem solutions, when monopolizing their influence on research; hiring and distribution of public funds (see Paehlke, 2004 for an entrenched 'Cult of Incompetence'). However, they also have the opportunity to promote any of these fields further towards a modern society using appropriate tools. I suggest that modelling definitely requires an appropriate funding structure for assuring progress. Setting up such a culture and infrastructure requires a sophisticated and contributing leadership with a global vision.

Models allow bringing people and lobbies together and locating data gaps to be overcome and improved with subsequent fieldwork and modelling (Scott *et al.*, 2002). Models offer the great advantage to be constantly improved and fine-tuned. Also, predictive modelling, as presented here, offer a major contribution in order to obtain a Population Viability Analysis that takes spatial issues serious. I believe that this subject should receive more attention because it can address a key topic in management, populations and habitats, in pro-active terms and before unwanted situations occur, *e.g.* Huettmann *et al.* (2005).

Depending on the wildlife species, on the type of habitat and the human dimension, some problems are more important than others for advanced modelling. However, due to the complex situations of most models currently one cannot present an always valid cookbook approach to successful wildlife and habitat modelling projects.

From the modelling experience, it was found that successful modelling requires manifold skills rarely taught at universities and during marine wildlife education, yet. They go beyond pure marine wildlife, fisheries, statistical and computer skills. Many pitfalls and problems can occur during such applications, and few published experiences, rules or standards exist how to improve marine wildlife and conservation modelling projects, how to avoid errors and how to implement models eventually in the political and legal decision-making process addressing conservation and sustainability. More guidance is needed. It was found that often an old-fashioned institutional culture has to be overcome first, and then replaced with a new digital one that handles spatial and interdisciplinary models as well as all of the related issues. This can turn into a non-trivial task. Many political, strategic and diplomatic approaches are still required to deal with subjective, and often unprofessional, attitudes towards spatial modelling. Valuable lessons can be learned here from the Remote Sensing discipline for instance, which similarly went through a learning phase and has now reached maturation and general acceptance (Franklin, 2001; Gottschalk *et al.*, 2005).

One should emphasize that highly accurate (spatial) predictions should be the goal for modelling projects because a generalized inference, and testable hypothesis, can be brought forward for a quantitative assessment, and if necessary, model improvement. This new culture counters the old-fashioned believe that only field observations are valid and convincing for a generalized inference in biological disciplines. I believe that modelling should remain open-minded and consider alternatives. Findings can still depend on the nature of the modelling algorithm, *e.g.* when it comes to the selection of predictors and actual spatial predictions. Therefore, a competitive multi-model approach should be encouraged (Burnham and Anderson, 2002) and I suggest assessing and challenging the traditional approaches such as simple hypothesis testing with p-value thresholds, and hand-drawn distribution maps from experts and GLM models as the ultimate paradigms. Instead, and in times of great data availability and high technological tools, one should promote that model project repeats and duelling models are wanted as a form of true hypothesis tests towards science-based adaptive management (Walters, 1986), improved models and decision-making of public resources using the best science principle (Sarewitz *et al.*, 2000). Therefore, sound assessments of model accuracies are crucial, and it is suggested to fully support any of these approaches, including the collection and compilation of alternative assessment data and evidences.

Conclusions and Outlook

Due to the existence of free data sets, predictive models are maturing and prove to be of major value to management. Data availability is increasingly less of a problem, but data quality and resolution still needs to be steadily improved on a global scale. Biologically meaningful marine biodiversity and prey information such as high-quality benthos, plankton and fish databases collected with a consistent protocol are still needed. The data overkill of the future needs to be tamed with appropriate software tools (Huettmann, 2005).

Competing models are part of a scientific investigation using hypothesis; they are required and important to improve spatial models and eventually increase model trust. Once evaluated, many models still lack their implementation into policy and management, and it is suggested to quickly improve this situation on a global scale towards a new digital data and model culture of the oceans and beyond for the global village. The awareness, education and support for modelling needs to be further improved in the public, agencies and among scientists and lawmakers. Modelling should become a legal requirement when dealing with endangered wildlife and habitat.

Acknowledgements

I gratefully acknowledge the approaches brought forward by S. Levitus at NOAA who presented a great vision for the oceans and beyond, often still not reached on this scale in terrestrial systems and globally. Further, I would like to thank all data providers and all individuals and agencies who make their data available to research, to students and to the public, free of charge and online. I am further grateful to B. MacDonald and A.W. Diamond for introducing me into modelling, and J and S. Linke for initial support and discussions. Salford Systems kindly provided some of the modelling software. V. Reifenstein helped with the figures. The conversation with L. Strecker, EWHALE lab members, and support from B. Bluhm was highly appreciated. An NCEAS workshop on 'Alternative Modelling Techniques' allowed for a tremendous support and opportunity complementing my knowledge on modelling science. This is EWHALE publication number 23.

References

Bookhout T.A. 1994. Research and Management Techniques for Wildlife and Habitats. Wildlife Society, Bethesda, MD

Brown R.G.B. 1986. Revised Atlas of Eastern Canadian Seabirds. Canadian Wildlife Service, Ottawa

Brown R.G.B., D.N. Nettleship, P. Germain and C.E. Tull. 1975. Atlas of Eastern Canadian Seabirds. Canadian Wildlife Service, Ottawa

Brown D., H. Hines, S. Ferrier and K. McKay. 2000. Establishment of a biological information base for regional conservation planning in north-east New South Wales. NSW National Parks and Wildlife Service. Occ. Paper 26

Burger A.E. 1995. Marine distribution, abundance, and habitats of Marbled Murrelets in British Columbia. In: Ralph CJ, Hunt GL, Raphael MG, Piatt JF (eds.) Ecology and Conservation of the Marbled Murrelet, Gen. Tech. Rep. PSW-GTR-152. Pacific Southwest Research Station, Forest Service, U.S. Department of Agriculture, Albany:295-312.

Burnham K. and D.R. Anderson. 2002. Model selection and multi-model inference: a practical information-theory approach. Springer, New York.

Czech B. and P. Krausman. 2001. The Endangered Species Act: History, Conservation and Public Policy. Johns Hopkins University Press.

Esanu J.M. and P.F. Uhlir. 2004. Open Access and the Public Domain in Digital Data and Information for Science: Proceedings of an International Symposium. U.S. National Committee for CODATA, National Research Council.

Fenske J. 2003. Modelling Waterbird Colonies of the Great Slave Lake, Northwest Territories, Canada: A predictive GIS Model of Species Occurrenec of Arctic and Black Terns, and American White Pelicans. Unpublished M.Sc. Thesis, University of Potsdam, Germany.

Fielding A.H. and J.F. Bell. 1997. A review of methods for the assessment of prediction errors in conservation presence/absence models. Env. Cons.24:38-49.

Ford A. 1999. Modelling the Environment: An Introduction to System Dynamics Modelling of Environmental Systems. Island Press.

Franklin S.E. 2001. Remote Sensing for Sustainable Forest Management. Lewis Publishers, Boca Raton.

Gottschalk T., F. Huettmann and M. Ehlers. 2005. Thirty years of analysing and modelling avian habitat relationships using satellite imagery data: a review. Int. J. for Rem. Sens. 26:2631-2656.

Graham C.H., S. Ferrier, F. Huettmann, C. Moritz and A.T. Peterson. 2004. New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecol. and Evol. 19:497-503.

Huettmann F. 2005. Databases and science-based management in the context of wildlife and habitat: towards a certified ISO standard for objective decision-making for the global community by using the internet. J. of Wildl. Manage. 69:466-472.

Huettmann F. 2004. Towards a Marine Environmental Information System (MEnvIS) for the Northwest Atlantic: Experiences and suggestions from a multi-disciplinary GIS conservation research project using public large scale monitoring and research data bases from the internet. 4th WSEAS (World Scientific and Engineering Academy and Society) Proceedings, Tenerife/Spain, December 2003.

Huettmann F. 2000a. Making use of public large-scale environmental databases from the WWW and a GIS for georeferenced prediction modelling: A research application using Generalized Linear Models, Classification and Regression Trees and Neural Networks. In: Tochtermann K and Riekert W.-F. (Eds.) "Hypermedia im Umweltschutz" Proceedings of Deutsche Gesellschaft für Informatik (GI) and Forschungsinstitut für anwendungsorientierte Wissensverarbeitung (FAW) Ulm. Umwelt-Informatik aktuell; Bd.24, Metropolis Verlag/Marburg. pp. 308-312.

Huettmann F. 2000b. Seabirds in the Marine Wilderness of the western North Atlantic. Sixth World Wilderness Congress Proceedings, Bangalore/India, Vol. II, Proc. RMRS-P-000. Ogden, UT, US Department of Agriculture, Forest Service, Rocky Mountain Research Station.

Huettmann F. and A.W. Diamond. 2001 Seabird colony locations and environmental determination of seabird distribution: A spatially explicit seabird breeding model in the Northwest Atlantic. Ecol. Mod. 141:261-298.

Huettmann F. and J. Linke. 2003. An automated method to derive habitat preferences of wildlife in GIS and telemetry studies: A flexible software tool and examples of its application. European J. of Wildl. Res. 49:219-232.

Huettmann F. and A.R. Lock. 1997. A new software system for the PIROP database; data flow and an approach for the seabird-depth analysis. ICES J. for Mar. Science, 54:518-523.

Huettmann F., S.E. Franklin and G.B. Stenhouse. 2005. Developing Future Landscape Scenarios for Grizzly Bear Habitat. Can. Forestry Chronicle 81:1-13

Huettmann F., D. Lank, E. Cam, R. Bradley, L. Tranquilla-McFarlane, L. Lougheed, C. Lougheed, P. Yen, Y Zharikov, N. Parker and F. Cooke (in review). Nest habitat selection of Marbled Murrelets (*Brachyramphus marmoratus*) in a fragmented Old Forest Landscape. Journal of Wildlife Management.

Levitus S. 1994. World Ocean Atlas 1994, 4 volumes, Washington, D.C., National Oceanic and Atmospheric Administration, National Oceanographical Data Center.

Malakoff D. 2003. Scientists counting on census to reveal marine biodiversity. Science. 302:773.

Manly F.J., L.L. McDonald, D.L. Thomas, L.T. McDonald, W.P. Erickson. 2002. Resource Selection by Animals. Kluwer Academic Publishers.

Nakazawa Y, A.T. Peterson, E. Martinez-Meyer and A.G. Navarro-Siguenza. 2004. Season niches of Nearctic-Neotropical migratory birds: implications for the evolution of migration. Auk 121:610-618.

National Oceanic and Atmospheric Administration NOAA. 2000. Coastwatch: West Coast Regional Node, Pacific Grove, CA. http://cwatchwc.ucsd.edu/cwatch.html. Accessed May 2002

Paehlke R. 2004. Democracy's Dilemma: Environment, Social Equity, and the Global Economy. MIT Press, Cambridge Massachusetts.

Pearce J.L. and S. Ferrier. 2000. Evaluating the predictive performance of habitat models developed using logistic regression. Ecol. Mod. 133:225-245.

Primack R.B. 1998. Essentials of Conservation Biology. 2. Edition. Sinauer Associates Publishers.

Rozwadowski H.M. 2002. The sea knows no boundaries: a century of marine science under ICES. University of Washington Press, University of Columbia Press.

Sarewitz D.R., A. Pielke and R. Byerly. 2000. Prediction: Science, Decision Making and the Future of Nature. Island Press.

Scott M.J., P.J. Heglund, M.L. Morrison, J.B. Haufler, M.G. Raphael, W.A. Wall and S.A. Samson. 2002. Predicting Species Occurrences: Issues of Accuracy and Scale. Island Press, Washington.

Shenk T.M. and A.B. Franklin. 2001. Modelling in Natural Management: Development, Interpretation, and Application. Island Press. Washington.

Sirois J., M.A. Kay and M.A. Fournier. 1995. The colonial waterbirds of Great Slave Lake, Northwest Territories: an annotated atlas. Occasioanl Papers, Canadian Wildlife Service No. 89. Ottawa.

Valavanis V. 2002. GIS Applications in Oceanography and Fisheries. Taylor and Francis.

Walters C. 1986. Adaptive Management of Renewable Resources. Blackburn Press, Caldwell New Jersey.

Ydenberg R.C. 1998. Evaluating models of departure strategies in alcids. Auk 115:800-801.

Yen P., F. Huettmann and F. Cooke. 2004. Modelling abundance and distribution of Marbled Murrelets (*Brachyramphus marmoratus*) using GIS, marine data and advanced multivariate statistics. Ecol. Mod. 171:395-413.

Zhang Y. and J.F. Grassle. 2003. A portal for the Ocean Biogeographic Information System. Oceanologica Acta. 25 (5):193-197.

The BODC Taxon Browser – A powerful search tool for the discovery of taxonomic information

Michael Hughes and Roy Lowry

British Oceanographic Data Centre
Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, United Kingdom.
E-mail: mhug@bodc.ac.uk

Abstract

The BODC Taxon Browser is used for the discovery of descriptive information about the biological taxa that are held in the archives of the British Oceanographic Data Centre (BODC). The main aspect of its functionality is the capability of hierarchical searching. This means that on entering a taxonomic search term, the browser will retrieve information for the term and information for all organisms taxonomically related to the term. The framework of the hierarchical search is derived from the Integrated Taxonomic Information System (ITIS). Other features include: automatic checking for synonyms where taxonomic names are not valid; search by ITIS code, BODC code, scientific name and common name; filter options to focus the search. This search tool is powerful – it goes beyond simply matching the exact term that is entered and so discovers information which users may not even realise they are looking for!

Keywords: Taxonomy; Hierarchical searching; Metadata discovery; Parameter dictionary; The Integrated Taxonomic Information System.

Introduction

The BODC Taxon Browser is a web-based search tool which enables easy navigation through the substantial amount of biological data which is managed by BODC. BODC holds nearly 400,000 biological records spanning 4,500 different taxa. Measurements range from the biomass of microscopic algae in the Indian Ocean to counts of Cetaceans in the North-East Atlantic. Each record is linked to detailed metadata (*e.g.* **what** was measured, **how** it was measured, **who** measured it, **where** and **when** it was measured) via a relational database management system (RDBMS). One branch of the RDBMS is a parameter dictionary which contains the "what" and "how" metadata. The BODC Taxon Browser allows the user to efficiently access this metadata for every taxonomic entity that is described within this parameter dictionary.

The parameter dictionary is constantly expanding and currently describes nearly 17,000 types of variables, 11,000 of which originate from taxonomic samples. On entering a taxonomic search term, the browser will retrieve metadata for the term and the metadata for all organisms taxonomically related to the term. So, if you search for "birds" you'll also get "penguins". This works even if the actual search term is not present in the dictionary. For example, a search for "*Cetacea*" (an Order of marine mammals) will return metadata for every genus and species of dolphin that is held in the dictionary even though the word "*Cetacea*" does not appear in any part of the dictionary. This is particularly helpful if users have only a general idea of the taxonomic information they are looking for.

Where possible, every taxonomic entity in the parameter dictionary is mapped to an entry in the Integrated Taxonomic Information System (ITIS). This gives credibility to the taxonomic information at BODC as ITIS is an authoritative taxonomic resource and it also enables the Taxon Browser to use the ITIS taxonomic hierarchy as a framework for the "intelligent" searching described above.

Other Taxon Browser features built on the ITIS framework are:

- Automatic checking for synonyms where taxonomic names are no longer considered to be valid
- Searching by common names.

To make the Taxon Browser functional, a number of stages of development were required, namely:

- Downloading a local copy of the complete ITIS database
- Mapping all BODC taxonomic entities to their equivalent taxa in ITIS
- Generating a taxonomic hierarchy from the ITIS tables
- Creating a web-based search interface to enable dynamic interaction between the user and the database.

This paper will outline these aspects of development and give some working examples of using the Taxon Browser.

Incorporating ITIS

ITIS provides reliable information on species names and their hierarchical classification. The database is regularly reviewed to add newly described species and to keep up-to-date with the validity of taxonomic classifications. Every scientific name in ITIS is accompanied by the author and date, taxonomic rank, associated synonyms and vernacular names where appropriate, data source information, data quality indicators and a unique taxonomic serial number (TSN). The TSN becomes the label for what is known as a "Taxonomic Unit". In order to integrate the BODC taxonomic data with the ITIS taxonomic units, every taxonomic entity at BODC is assigned a TSN where available (see section 1.2).

Downloading ITIS

The ITIS database is updated on a monthly basis and the latest version can be freely downloaded from: ftp://ftp-fc.sc.egov.usda.gov/ITIS/. The folder at this location contains text-only versions of all the tables used in the ITIS database and an SQL file with the necessary code to set up the tables locally. The ITIS tables were downloaded according to the instructions at: http://www.itis.usda.gov/ftp_download.html. This process is repeated every two months to keep up with the latest updates to the ITIS database.

The ITIS table of relevance to the Taxon Browser are:

- ITIS Taxonomic Units – includes scientific names, usage information, hierarchy information, links to references, credibility information
- ITIS Vernaculars – common names in various languages which are linked via a TSN to their relevant taxonomic units
- ITIS Synonym Links – where taxonomic name usage is considered invalid (for animals and bacteria) or not accepted (for plants and fungi), the valid or accepted alternative is contained here.

The BODC-ITIS map

Once local copies of the ITIS tables were generated, the scientific taxon names at BODC could be automatically mapped to the ITIS taxonomic units. To do this, a series of SQL statements in Oracle were used to identify matches between the BODC scientific names and the ITIS scientific names. The relevant ITIS TSN was then appended to all BODC parameter dictionary entries describing sampling events for that taxon.

Occasionally, automatic mapping was not successful and records needed to be manually checked. The two main reasons for this were:

- Spellings differed between BODC and ITIS (usually due to spelling mistakes)
- A scientific name did not appear in ITIS.

Where spellings differed slightly (*e.g. Chaetoceros pelagicum* vs. *Chaetoceros pelagicus*), the BODC name was altered to comply with ITIS. Where scientific names appeared in BODC but not in ITIS, the names were submitted to ITIS according to the guidelines at http://www.itis.usda.gov/submit_guidlines.html. The web resources used to find information for the submission of species names to ITIS were: Algaebase (Guiry and Nic Dhonncha, 2005), The Ciliate Resource Archive (Lynn, 2003), The World of Protozoa, Rotifera, Nematoda and Oligochaeta (Inamori, 2003) and the Check-list of Turkish Seas Microplankton (Koray *et al*, 1999).

The BODC-ITIS map fulfils a number of roles:

- It provides a link to ITIS taxonomic units, common names, synonyms and any other useful information provided by ITIS
- The columns of the table are arranged into "semantic elements" which are the building blocks that are used to automatically generate descriptive full titles in the parameter dictionary.

An example of a parameter dictionary full title is:

Carbon biomass of Bacillariophyta (ITIS 2286) centric 30um per unit volume of the water column by optical microscopy and abundance to carbon conversion by unspecified algorithm

The semantic elements for this title are:

- Parameter: Carbon biomass
- Taxon name: Bacillariophyta
- Taxon code: 2286
- Taxon class: centric 30um
- Parameter compartment: per unit volume of the
- Compartment: water column
- Compartment class: not specified (hidden)
- Sample preparation: not specified (hidden)
- Analysis: optical microscopy
- Data processing: abundance to carbon conversion by unspecified algorithm

The full title was created by the concatenation of the fields in the BODC-ITIS map to produce a humanly readable sentence. The fields of the BODC-ITIS map contain the semantic elements. Every full title generated has the same arrangement of semantic elements to maintain consistency and accuracy when creating parameter definitions. Each semantic element is part of a controlled vocabulary meaning that there are a limited number of options for the words that can be used for any particular semantic element. For example, “Parameter” can have values including, “Carbon biomass”, “Abundance” and “Count”; “Taxon name” must be the same as the name published in ITIS where available; “Compartment” can be “bed”, “sediment”, “water column” or “suspended particulate matter”. The elements are joined by linking words such as “of” or “in the”. This system permits a rich vocabulary of definitions for the description of the many forms of data at BODC and is also easily machine readable.

Building the taxon tree

After populating the BODC-ITIS map, a program was written to generate a hierarchy of all the ITIS taxonomic units. This forms the framework of the Taxon Browser’s hierarchical search. It enables the Taxon Browser to navigate the ITIS hierarchy and extract every taxonomic entry in the BODC dictionary at the same taxonomic level as the search term and also all the related taxa at lower levels.

The hierarchy was built using a field of the ITIS Taxonomic Units table called “Parent TSN”. This is a direct link between every ITIS taxonomic unit and the taxonomic unit directly above it in the ITIS taxonomic hierarchy. For example, the species *Chaetoceros pelagicus* has a parent TSN of “2758”. The scientific name with a TSN of 2758 is the genus *Chaetoceros*. *Chaetoceros* has a parent TSN of “572759” which is the TSN for Family Chaetocerotaceae. This path can be followed all the way to the kingdom level.The representation of the ITIS taxonomic hierarchy that is generated by the tree-building program consists of a source TSN followed by a string of 216 bits (see figure 1). Every 8 character section of the string represents a single taxonomic level. The default setting of the hierarchy string is 216 x's. This equates to 8 x's per taxonomic level. A string is populated by placing the source TSN in the position on the hierarchy string appropriate to its taxonomic level. For example, the TSN for a genus name is placed at character positions 137-144 in the hierarchy string; the TSN for a species name is placed at positions 169-176. When the source TSN is in place, every parent TSN for this is inserted into the string in the appropriate place. A separate hierarchy string is created for every taxonomic unit in the ITIS database.

Source TSN: 2814
Hierarchy:
xx202422xx590735xxxx2286xxxxxxxxxxxxxxxxxx590736xxxxxxxxxxxxxxxxxxxxx
xxxxx590748xxxxxxxxxxxxxxxxxxxxxxxxxx572759xxxxxxxxxxxxxxxxxxxxxxxxxx
xx2758xxxxxxxxxxxxxxxxxxxxxxxxxxxx2814xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
xxxxxxxxx

Fig. 1. The hierarchy string for species Chaetoceros pelagicus. *The species portion of the hierarchy string is represented by "xxxx2814". Its taxonomic parent is represented by "xxxx2758", which is the TSN for genus* Chaetoceros. *Continuing a walk up the string identifies every parent taxon for* Chaetoceros pelagicus *up to the kingdom level (202422, kingdom Plantae). Every empty taxonomic position is left as 8 x's to maintain the length of the hierarchy string. This becomes useful later for checking for TSNs in specific taxonomic positions. For example, in between 572759 and 2758 are unoccupied taxonomic levels that represent subfamily, tribe and subtribe; after 2814 are unoccupied taxonomic levels that represent subspecies, variety, subvariety, form and subform.*

Continuing the example from figure 1, there are many other species of *Chaetoceros*, all having identical hierarchy strings from the kingdom to the genus level, *i.e.* from character positions 1 to 144. This means that every hierarchy string with "xxxx2758" at the genus level is exactly the same from positions 1 to 144 and also represents a hierarchy for a species, subspecies, variety, subvariety, form or subform of the genus *Chaetoceros*. Selecting the source TSN for all these partially matching hierarchies allows the Taxon Browser to retrieve all the relevant parameter dictionary records regarding members of the *Chaetoceros* genus.

From a wider perspective, any hierarchy string that has a TSN ending at character position *n* shares the same hierarchy from positions 1 to *n* as any other hierarchy string that has the same TSN ending at character position *n*. A hierarchy string can be sliced at the boundary of any 8-character section to widen or narrow the range of source TSNs that are extracted by the Taxon Browser. This allows the user to enter a search term at any taxonomic level and retrieve all of its relatives. A search at the kingdom level, for example will extract all taxa from BODC that have matching hierarchy strings for positions 1 to 8 for any particular kingdom. It is possible to type in "Animalia" and retrieve every taxon at every available taxonomic level for entities considered to be animals even though there is no entry in the parameter dictionary that is explicitly described as being an animal. Similar searches can be conducted for "Aves" or "Birds" or even "Oiseaux" or perhaps the user would like to narrow the search slightly and look for "Gulls". Again, these are all words that do not appear anywhere in the parameter dictionary but will still yield comprehensive results.

This method works for searching all taxa that occur in the ITIS database. However, a special case arises when a taxon is considered "invalid" or "not accepted" by ITIS. For these cases, no parent TSN is provided in ITIS Taxonomic units and so there is no basis for the tree builder to generate a hierarchy string. For the hierarchical search to be fully functional, the taxonomic hierarchy must be derived from the "accepted" or "valid" synonym. Using the ITIS Synonym Links table, it is possible to find the valid TSN and superimpose its hierarchy onto the invalid TSN. An alternative strategy could be to change all invalid names at BODC into their valid synonyms but it is better to alter the data coming into BODC as little as possible. Using the ITIS Synonym Links table means that this can be avoided and the user can be informed that a particular taxonomic name is not valid and of its valid usage.

Another special case that arises from an invalid synonym being assigned the parental hierarchy of its valid counterpart is when the adopted parent is at the same taxonomic level as the invalid name. For example, **class** Solenogastres is invalid. Its valid synonym is **subclass** Chaetodermomorpha. The parent taxon for both these names is **class** Aplacophora. Since a class can not have a taxonomic parent which is also a class, the original TSN is loaded into the hierarchy string in the position that the valid synonym occupies. The program that builds the taxon tree table takes all of these issues into account and automatically adapts to the special cases.

Using the Taxon Browser

This section concerns the client-side of the Taxon Browser and describes the user interface and the various search options and features that the browser provides.

User interface

The user interface (see figure 2) was written in Perl and uses the Common Gateway Interface (CGI). CGI enables interaction between a client and a server via the World Wide Web. The program is embedded with HTML to provide the front-end graphics on the Web and SQL to fetch information from the Oracle database at BODC. The user can select different searching methods (see section 2.2), apply options to filter the search (see section 2.3) and search within BODC or ITIS.

Search options

Search by scientific name:

- The Taxon Browser matches the scientific name in ITIS Taxonomic Units and extracts its TSN
- The taxonomic hierarchy is scanned to select every source TSN with a hierarchy string that includes the search TSN

- Every TSN selected is added to an array
- The parameter dictionary metadata for each TSN in the array is extracted and displayed to the user (see figure 3).

Search by common name:

- ITIS Vernaculars is scanned for the TSN belonging to the matching common name
- The search follows the last two steps of "Search by scientific name".

Search by BODC code:

- BODC-ITIS map is scanned for the TSN belonging to the matching BODC parameter code (a unique identifier for each entry in the parameter dictionary)
- The search follows last two steps of "Search by scientific name".

Search by ITIS TSN

- The search follows last two steps of "Search by scientific name".

Features

Automatic synonym conversion:

- Where a search term is invalid or not accepted, the Taxon Browser takes the alternative accepted TSN from ITIS Synonym Links
- The taxonomic hierarchy is scanned to select every source TSN with a hierarchy string that includes the original search TSN or the accepted TSN
- The search follows steps 2-3 of "Search by scientific name".

Filter by parameter type

- The user can adapt the search to select only those dictionary records that deal with abundance or biomass data, for example.

Hierarchical / non-hierarchical searching

- The user can select whether to perform a full hierarchical search as described previously or to retrieve only the parameter dictionary metadata for the actual search term.

Advanced options

- The user can customise the items of parameter dictionary metadata to be displayed by the Taxon Browser. For example, the user may be interested in how samples where analysed but not interested in how the data was processed and can change the settings accordingly.

Search ITIS

- For a direct link to the complete taxonomic information in ITIS, a search can be directed to the ITIS website.

Fig. 2. The Taxon Browser search interface.

SEARCH RESULTS for "chaetoceros":

BODC code	ITIS tsn	Scientific Name	Taxon Class	Parameter	Compartment	Rank	Usage
P030M00L	2758	Chaetoceros	auxospore	Abundance	water column	Genus	accepted
P030M00K	2758	Chaetoceros	large size	Abundance	water column	Genus	accepted
P030M00E	2758	Chaetoceros	medium size	Abundance	water column	Genus	accepted
P030M00Z	2758	Chaetoceros	not specified	Abundance	water column	Genus	accepted
D030M00C	2758	Chaetoceros	resting spores	Abundance	sediment	Genus	accepted
■ ■ ■							
C030M76Z	550512	Chaetoceros rostratus	not specified	Carbon biomass	water column	Species	accepted
P030M91Z	573625	Chaetoceros volans	not specified	Abundance	water column	Species	not accepted
C030M91Z	573625	Chaetoceros volans	not specified	Carbon biomass	water column	Species	not accepted
P030M07Z	610085	Chaetoceros bulbosum	not specified	Abundance	water column	Species	accepted
C030M07Z	610085	Chaetoceros bulbosum	not specified	Carbon biomass	water column	Species	accepted
P030M02A	2771	Chaetoceros atlanticum neapolitanum	not specified	Abundance	water column	Variety	accepted
C030M02A	2771	Chaetoceros atlanticum var. neapolitanum	not specified	Carbon biomass	water column	Variety	accepted
I3A97A37	2772	Chaetoceros atlanticum skeleton	not specified	Abundance	water column	Variety	accepted
P030M54A	2845	Chaetoceros simplex calcitrans	not specified	Abundance	water column	Variety	accepted
C030M54A	2845	Chaetoceros simplex calcitrans	not specified	Carbon biomass	water column	Variety	accepted

Rows selected: 157

Fig. 3. A selection of metadata from the BODC parameter dictionary retrieved by searching for scientific name "Chaetoceros". Each column forms the semantic elements used to define a parameter.

Discussion

Taxonomic awareness

Within BODC, the linking of taxonomic entities to ITIS has forced us to become more aware of official taxonomic naming conventions and encouraged the production of a protocol for creating new taxonomic entries for the parameter dictionary. Part of this protocol is that every species name added to the parameter dictionary must comply with ITIS and must also be given an ITIS TSN wherever possible. This practice has dramatically increased the quality and credibility of the parameter dictionary definitions concerning taxonomic entities.

Semantic modelling

The semantic model was a breakthrough for the automatic mapping of BODC to ITIS. The laborious task of manually extracting and matching taxonomic names from previously semantically uncontrolled parameter definitions became an easy task via a simple SQL statement. The semantic model also has the potential to become the basis to create a web service for sharing taxonomic parameter information across the internet.

Taxonomic standards

A large part of the task when setting up the Taxon Browser was to submit species names to ITIS. This required extensive searching on the internet for accurate references and authorship information. It also highlighted the need for standardisation of scientific names within the taxonomic community. For example, a search in Google for "*Chaetoceros compressus*" gives 164 hits. A search for "*Chaetoceros compressum*" gives 119 hits. How does a non-expert user decide which name to use? Taking the name with the most hits seems an obvious strategy but this is by no means accurate. If everyone did this, an incorrect name could rise to precedence by a sort of runaway selection whereby an incorrect name with a few more hits than a correct name would be quoted more frequently. Consequently, the gap in hit numbers between the two names would widen, causing users to choose the incorrect name over the correct one with an increasingly high frequency.

A single comprehensive database that incorporates taxonomic information from a distributed community of experts could solve this type of problem. ITIS is a big step towards this but lacks the resources to keep up with the rate at which species names are submitted. Submissions made by BODC at the end of 2003 have still not made it into the ITIS database. With limited resources, ITIS must also focus on certain species groups meaning that some are not yet included in the database. For example, only two species of *Strombidium* (a marine ciliate) are listed in ITIS whereas there are 74 in the European Register of Marine species (ERMS, http://www.marbef.org/data/erms.php). The Taxon Browser encounters a problem here because BODC has 39 species of *Strombidium* which are not in ITIS and so have not been included in the taxonomic tree that the Taxon Browser searches through. A user of the Taxon Browser will not see that there is information on these *Strombidium* species at BODC unless they use the non-hierarchical search option.

It is a huge task to collate the names of every single described species and organise them into ever-changing taxonomic hierarchies. It is inevitable that different expert opinions will arise of where a species should fit into a hierarchy or what it should be called and perhaps this is beyond the scope of a single taxonomic information centre. An alternative to a single central database is a number of expert databases with their own particular focus on selected groups of organisms. There are a number of these available but uncertainty can arise when two databases disagree. For example, ITIS considers *Emiliania huxleyi* to be a not accepted synonym for *Coccolithus huxleyi*. However, ERMS accepts the name *Emiliania huxleyi* and gives it a non-accepted synonym of *Pontosphaera huxleyi* with no mention of *Coccolithus huxleyi*. So, which information source should be used? A search in Google gives 16,700 hits for *E. huxleyi* but only 168 hits for *C. huxleyi*. This shows that the general consensus is to use *E. huxleyi* but there must be a reason why this name appears as not accepted in ITIS. Having many separate on-line databases also makes searching less efficient as the user must search each database separately for taxonomic information. A useful tool would bridge these databases and search all of them in one go. The uBio Name Mapper (The Marine Biological Laboratory, 2004) works along these lines but does not cover a wide enough range of databases.

Conclusion

The BODC Taxonomic Browser has the potential to be a very useful tool for the discovery of data in the BODC archives. Some work is required to make it fully operational but when it is released on the web, it is expected that people from around the world will access it to view the types of taxonomic data at BODC. The mapping of names to ITIS was a very useful exercise in improving the BODC parameter definitions but it also highlighted the fact that BODC should be aware of a world of taxonomic databases beyond ITIS and consider how these could be integrated.

Acknowledgements

Thanks to: The ITIS team for producing a very valuable database and providing comprehensive documentation and support to get full use out of it. Steve Loch at BODC for his ideas on generating the hierarchy strings. Richard Downer at BODC for his advice on the environmental settings for using Perl CGI on the BODC computer system. Gwen Moncoiffé at BODC, Toby Tyrell at the Southampton Oceanography Centre, Sonia Batten at Plymouth Marine Laboratory, Jeremy Young at the Natural History Museum and Mike Guiry at the National University of Ireland, Galway for taxonomy advice during the manual mapping of BODC terms to ITIS.

References

Guiry M.D. and E. Nic Dhonncha. 2005. *AlgaeBase version 2.1*. World-wide electronic publication, National University of Ireland, Galway. Available online at http://www.algaebase.org. Consulted on 25 January 2005.

Inamori Y. 2003. The World of Protozoa, Rotifera, Nematoda and Oligochaeta. National Institute for Environmental Studies, Japan Environmental Agency. Available online at http://www.nies.go.jp/ chiiki1/protoz/. Consulted on 1 December 2003.

Koray T., S. Gokpinar, L. Yurga, M. Turkoglu and S. Polat. 1999. Microplankton species of Turkish Seas. Available online at http://bornova.ege.edu.tr/~korayt/plankweb/chklists.html. Consulted on 1 December 2003.

Lynn D.H. 2003. The Ciliate Resource Archive. Available online at http://www.uoguelph.ca/~ciliates. Consulted on 1 December 2003.

Costello M.J., P. Bouchet, G. Boxshall, C. Emblow and E. Vanden Berghe. 2004. European Register of Marine Species. Available online at http://www.marbef.org/data/erms.php. Consulted on 26 January 2005.

The Integrated Taxonomic Information System on-line database. Available online at http://www.itis.usda.gov. Consulted on 9 December 2004.

The Marine Biological Laboratory, Massachusetts, USA. 2004. The Universal Biological Indexer and Organizer (uBio) Name Mapper Available online at http://uio.mbl.edu/services/pleary_working/ treeserve.php. Consulted on 26 January 2005.

Linear referencing as a tool for analyses of organic material deposition along a sandy beach of Gdansk – Sopot - Gdynia (Polish coast of Baltic Sea)

Monika Kędra[1] and Jacek Urbański[2]

[1]Institute of Oceanology PAS, Sopot 81 – 712, Powstańców Warszawy 55, Poland
E-mail: kedra@iopan.gda.pl

[2]Institute of Oceanography, University of Gdańsk, Gdynia 81-378, Al. Marszałka Piłsudskiego 45, Poland

Abstract

Linear referencing is a technique provided by GIS. It supplies tools that help in dealing with data associated with any kind of linear features. In recent years dense algal mats covering the shoreline and algal blooms in the Baltic region became a serious problem and create much concern. The linear referencing technique was used to analyse the dynamic process of organic material deposition along the shoreline. Sampling was done nine times along the sandy beach of Gdańsk – Sopot – Gdynia. The total amount of algal debris washed ashore varied a lot each time and on average was 252.64 tons ±220.97 tons. The linear referencing technique proved to serve well in dealing with processes occurring along any feature that may/can be treated as linear.

Keywords: Linear referencing; organic material deposition; Baltic Sea.

Introduction

Linear referencing is technique provided by GIS. Previously linear referencing was mainly used in GIS - transport sciences for management and for querying spatially and temporally referenced transportation data. In case of transportation linear referencing is a core method due to the fact that transportation features are linear in nature (Sutton and Wyman, 2000).

Linear referencing supplies tools that help in dealing with data associated with any kind of linear features. It reduces the effort of maintaining, organising, analysing and controlling any data describing processes that occur along existing linear features. It allows associating any multiple a set of objects with linear features. Then, it enables querying, editing or analysing attributive data sets without affecting the linear feature (Brennan, 2002).

Recently some effort was made to introduce linear referencing to environmental sciences. The technique was used to investigate coastal erosion and river systems classification.

In our case deposition of organic material along the sandy shoreline was analysed with the linear referencing technique. Dense algal mats covering the shoreline and algal blooms in the Baltic region became/are becoming a serious problem and created/creates much concern (Kotwicki *et al.*, 2002). Sandy beaches are very important due to their economical and social values. They also play an important role in the ecosystem as active biological filters. High organic material deposition along the shoreline strongly affects their socio-economical values and may influence the biota associated with interstitial system that plays a great role in processing organic material and returning nutrients back to the sea. The presence of green algal mats in shallow soft bottom areas could cause large changes in the local ecology (Isaksson and Pihl, 1992).

It is essential to assess the amount and volume of organic material and then to analyse biological processes connected with algae debris along the shoreline and linear referencing may serve this aim well.

Material and methods

The material was collected along the sandy beach of Gdańsk – Sopot – Gdynia (Polish coast of Baltic Sea) (fig. 1). There were 9 sampling campaigns from June to August 2004, carried out every week. Geographic position of each piece of organic material washed ashore was measured by using a GPS and portable GIS system. At the same time the volume of organic debris was measured. Small amounts were assessed as 1 dm^3. In case of larger amounts, length, width and depth of each assembly were measured and then the volume was calculated. Also the average volume of organic material deposited on every meter of the beach was calculated.

In Sopot additional sampling was made. 1dm^3 of swash wash water was taken in June and the density of drifting algae was assessed. Comparing these findings with the amount found on the beach, the percentage of the algae found in the water and that wash ashore was determined.

On average 1 dm^3 of algal debris deposited on the beach weighted 0.88kg (±0.10). These results were used to estimate the total amount of organic material found on the beach in comparison to the total volume counted from the sampling carried out along the Gdańsk – Sopot – Gdynia shoreline.

By employing the linear referencing method, geographic data may be stored by using a relative position along an already existing linear feature (route). A route is a geographic feature that is represented by a line, with a unique identifier and is stored with/in the geometry measurement system. In this case a shoreline from Gdansk to Gdynia is treated as one route. Any geographic feature occurring along the route should be treated as a route event. Every volume of organic debris is treated as a line event located along the route (shoreline). Linear referencing allows analysing dynamically changing objects with relative position on linear features. The objects are sampled in two dimensions, using x, y coordinates.

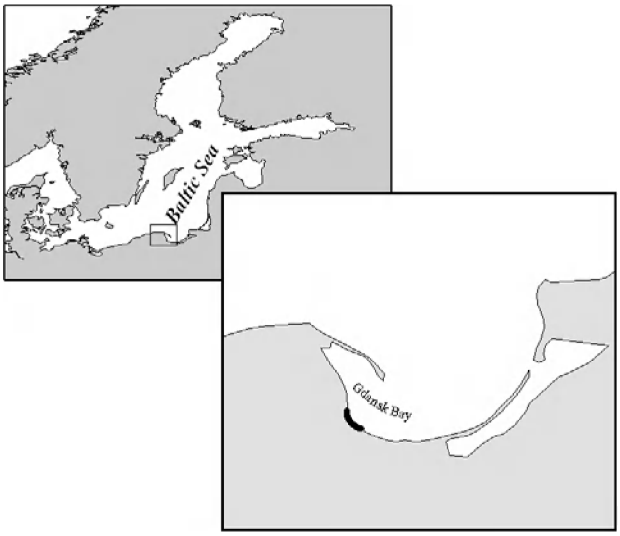


Fig. 1. Sampling area – Gdańsk – Sopot – Gdynia; Polish coast of Baltic Sea.

Results and conclusions

The volume and the amount of algae debris found along the coast varied the most in July. The highest amount was observed on 14.07.2004 while the lowest on 29.07.2004 (fig. 2).

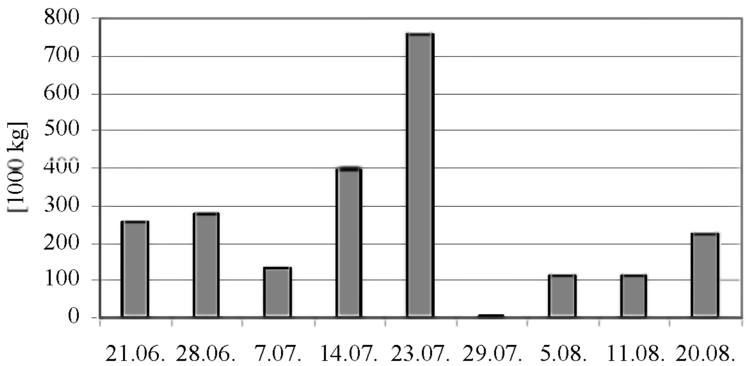


Fig. 2. Total amount of algae debris found along the coast in summer 2004.

The mean volume of algae debris per meter beach varied the most in July (fig. 3). All the 9 samplings carried out through the summer season proved that the manner in which the debris was deposited on the beach might not be treated as casual.

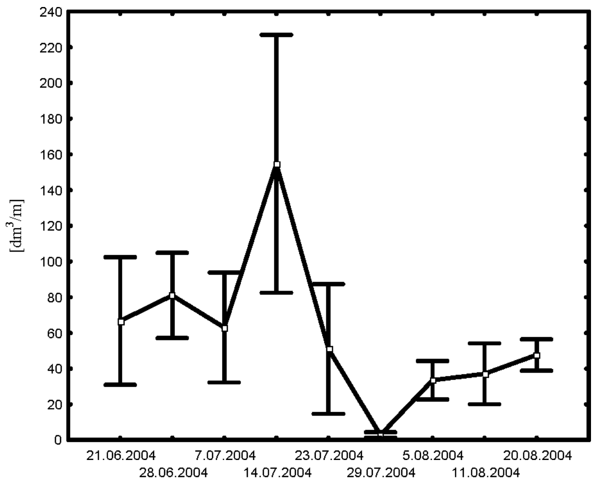


Fig. 3. Mean volume on every meter of the beach of algae debris found along the coast in summer 2004.

In June the mean density of algae in swash water was assessed as 13.5kg for each 1m^3. 64% of algae from the swash water were deposited on the beach.

Different natural phenomena occurring along the coastline have a linear and patchy character. For analysing such processes and for estimating parameters describing its temporal and spatial changes a precise knowledge regarding its occurrence is needed. That requires very dense and detailed point sampling or some linear framework might be an alternative. The example of organic material deposition along a sandy beach of Gdansk – Sopot – Gdynia showed that linear referencing technique may play such role. The linear referencing technique proved to serve good as a tool for storing, analysing and visualising environmental data connected with the shoreline.

Acknowledgements

We would like to thank Lech Kotwicki and Jan Marcin Węsławski for providing unpublished material.

References

Brennan P. 2002. Linear referencing in ArcGIS. GIS by ESRI. United states of America, Redlands. 165p.

Isaksson I. and L. Pihl. 1992. Structural changes in benthic macrovegetation and associated epibenthic faunal communities. Netherlands Journal of Sea Research, 30:131-140.

Kotwicki L., J. Danielewicz, M. Turzyński and J.M. Węsławski. 2002. Preliminary studies on the organic master deposition and particle filtration processes in a sandy beach in Sopot – Southern Baltic Sea. Oceanological Studies, 31(3-4):71-84.

Sutton J.C. and M.M. Wyman. 2000. Dynamic location: an iconic model to synchronize temporal and spatial transportation data. Transportation Research Part C 8:37-52.

Development of an updating information system on Decapoda Crustacea museum collections, useful in research and education

Koukouras A., M.-S. Kitsos, I. Kirmitzoglou and N. Chartosia

Department of Zoology, School of Biology, Aristoteleio University of Thessaloniki, GR-541 24 Thessaloniki, Greece.
E-mail: akoukour@bio.auth.gr

Abstract

Natural-history museum collections contain large amounts of historical and contemporary data for the assessment of global invertebrate biodiversity. Nevertheless, these data can not be easily accessed due to a series of constrains, including the lack of relative internet-based databases.

Towards this end, in the Zoological Museum of the Aristoteleio University of Thessaloniki (ZMAUTH) a new information system is under development on the Decapoda Crustacea collections. ZMAUTH hosts large collections of decapod species from all habitats of the Aegean Sea and other Mediterranean areas. Recently, all these decapod specimens along with information on the habitat from which they were collected were computerized and a dynamic database was created, also suitable for use through the internet. In this database, a complete bibliographic list of the decapod species known from the NE Atlantic and the Mediterranean Sea is also included as well as the relevant information on the habitat of each one of them.

Furthermore, all literature information concerning the systematics, zoogeography, ecology, biology, conservation and cultural status of every registered decapod species are being computerised in the database.

In this study, we demonstrate the procedures through which a user of this database can: (1) acquire the existing literature information for a species or a group of species, (2) acquire new scientific information, (3) assess, through a comparison his own relative data and (4) educate students and young researchers.

Keywords: museum collections; Decapoda; diversity databases.

Introduction

A significant proportion of the existing scientific information concerning global invertebrate diversity is in the form of museum collections. Specimens in these collections have been used to map species historical and present distribution and estimate species richness and diversity (Ponder *et al.*, 2001; O' Connell *et al.*, 2004; Suarez and Tsutsui, 2004).

However, the dissemination of the information stored in the museum collections is little and sparse due to the fact that access to collections and retrieval of specimens can be a difficult and lengthy process, constrained by work force, funding or both. Furthermore, despite some progress in the computerization of the collections, the resulting databases are not commonly available since most of them cannot be accessed through the internet (O' Connell *et al.*, 2004).

Towards these ends, the development and utilization of an information system on Decapoda Crustacea museum collections is presented in this study. This on-line database has resulted from the appropriate computerization of the numerous decapod specimens deposited in ZMAUTH along with literature information on the decapod species of the Mediterranean, the Black Sea and the NE Atlantic. The development of this information system intends to: (1) assemble and disseminate the existing information on the registered decapod species, (2) facilitate access to ZMAUTH museum collections, (3) provide tools for the acquisition of new scientific information based on the data resulting from the museum collections, (4) provide the appropriate web interface so that it can be integrated in the education of students in the relative disciplines.

Description of the information system

The design and construction of the information system is based on Microsoft Access interrelated tables. At the initial stage of the information system development a detailed list of all decapod species known from the Mediterranean, the Black Sea and the NE Atlantic, as given in the review of Udekem d'Acoz (1999), were registered in a MS Access table. For each species, along with the scientific name and authority, its given geographical distribution in the certain areas as well as information on its bathymetry and habitat type was also registered in the database.

At the same time, all museum specimens of the large decapod collections hosted in ZMAUTH were registered in an interrelated Access table of the database along with the sampling station data. Until now, a total of 235 decapod species have been registered in the database.

During the second stage of the information system development, which is still in progress, a thorough review of the international relevant literature is being carried out and all the information concerning the registered species is being gathered and then categorized in the following thematic sections: systematics, zoogeography, ecology, management and cultural status. All this information is registered in the database as literature citations. The dissemination of the available information is materialized through the web site of ZMAUTH (http://zoological-museum.bio.auth.gr), through which a possible user can access the information system on the decapod collections. The web site runs in a client-server mode and is built with the ASP (Active Server Pages) framework, permitting the dynamic content of the web pages.

A possible user starts navigating on the museum web site by selecting the invertebrate collections link in the main home page. In this way he is transferred to a web page containing illustrated links of the major invertebrate taxa represented in the museum's collections. By clicking the Arthropoda link he can navigate down to Decapoda through a series of static web pages, each one representing a lower rank in the systematic hierarchy of Arthropoda. The use of images and sounds in the design of these static web pages facilitates the use of this information system for educational purposes. Finally by selecting a certain decapod taxon the database is queried through SQL (Structured Query Language) and a list of the available species of this group, registered in the museum collections is given.

The selection of a certain species from this list leads the user to the species main web page containing the species scientific name and authority, an image of the species and button-links for the 6 main thematic sections (Systematics, Zoogeography, Ecology, Biology, Specimens, Management, Cultural) containing information about the certain species.

The Systematics web page contains the most important literature references concerning the etymology of the species name, major synonyms and distinct key characters for its identification as well as detailed images of these characters. Each of the references given has a specific code number which appears as a text-link next to the reference. By clicking on this link a new window pops up giving the reference full details. Links to other relevant internet resources are also provided.

The Zoogeography link in the species main page leads to a new dynamic web page where by using Macromedia Flash technology, a dynamic map is created demonstrating the distribution of the certain species in the Mediterranean regions, the Black Sea, the Red Sea and the NE Atlantic, according to the literature information stored in the database. These data are depicted on a raster map image which shows the certain geographical areas with two different colours according to the presence or absence of the species (Fig. 1).

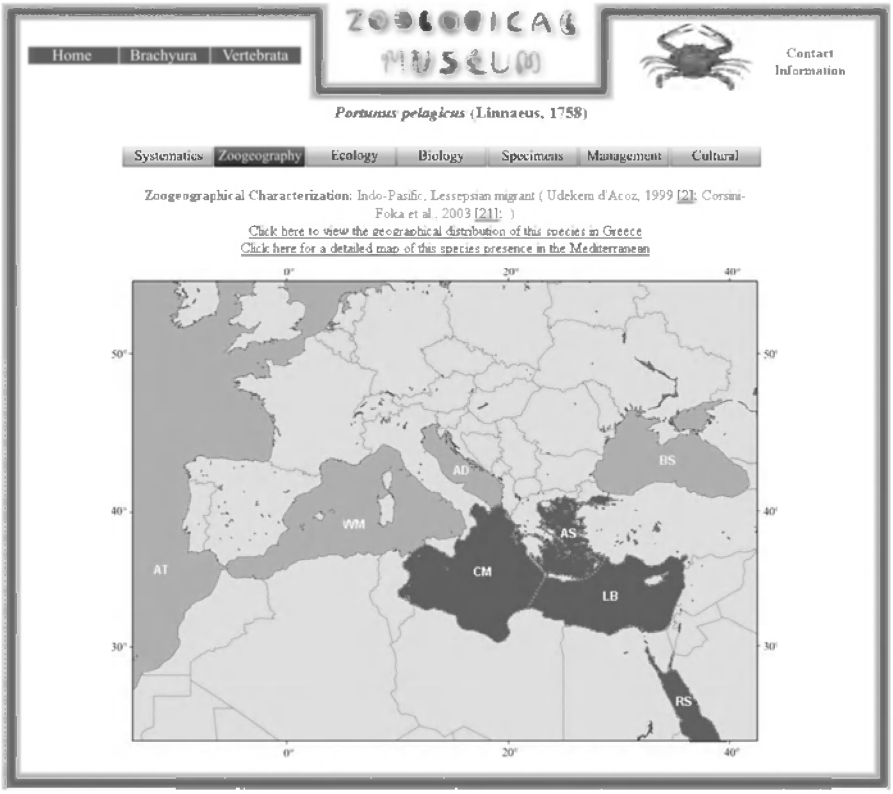


Fig. 1. The "Zoogeography" web page for the species Portunus pelagicus *(Linnaeus, 1758). Dynamic map indicating the distribution of this species in the certain Mediterranean areas, the Black Sea, the Atlantic and the Red Sea. Black colour represents the geographical areas this species has been found according to the literature.*

Following another link in the Zoogeography web page of a certain species, a dynamic map is created demonstrating the distribution of this species in the Aegean Sea as it results from the museum collection data (Fig. 2). Each dot in the map represents a locality where at least one registered specimen of the species has been collected.

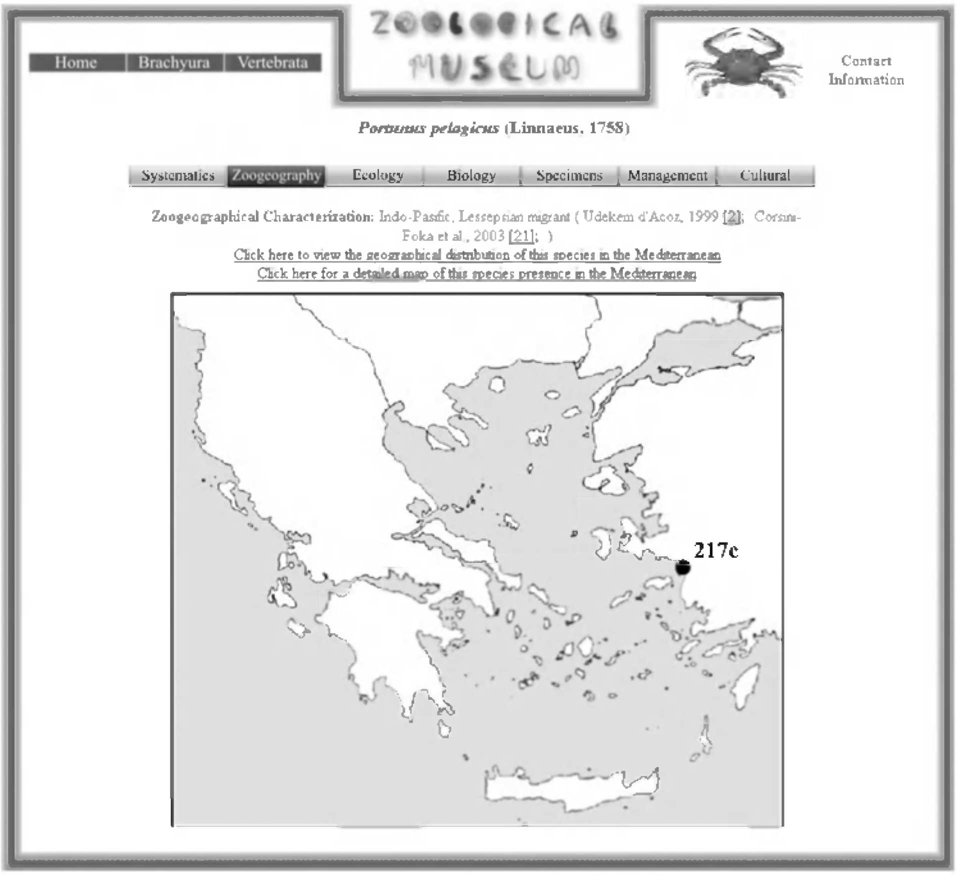


Fig. 2. The "Zoogeography" web page for the species Portunus pelagicus. Dynamic map indicating the distribution of this species in the Aegean according to the museum collections data. The black dot represents the station from which the only specimen of the collections has been collected. Bold number is the station code number.

Concerning the decapod species, which are listed as Lessepsian migrants, more detailed information is provided on their zoogeographical distribution. In the Zoogeography web page of each one of these species, a dynamic map of the Mediterranean is given, where all previous literature records of this species are shown as dots on the locality where the species has been reported from (Fig. 3). In the same map, the localities where the registered collection specimens of the certain Lessepsian species have been collected are also shown as dots of different colour (Fig. 3). Thus, a possible user can compare the geographical distribution of a given Lessepsian species as results from the literature and as it results based on the museum collection data and assess whether the latter extend the known geographical distribution of the species.

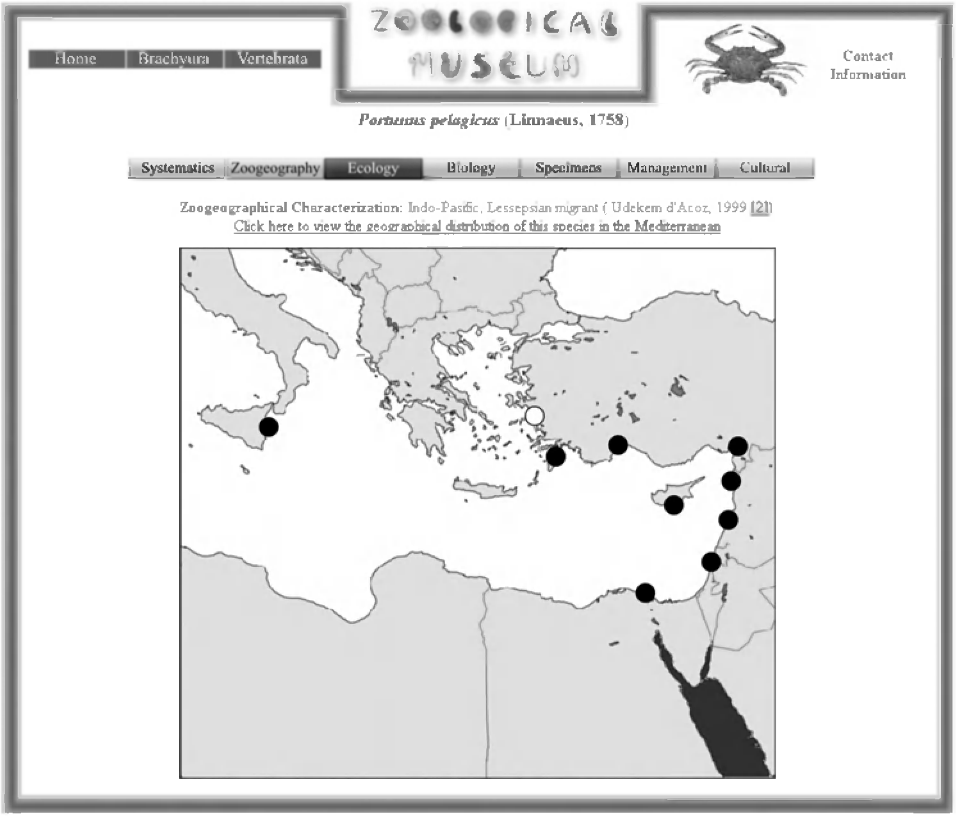


Fig. 3. The "Zoogeography" web page for the species Portunus pelagicus, *a Lessepsian migrant. Dynamic map of the Mediterranean indicating the species geographical distribution according to the literature and the museum collections data. Each black dot indicates a previous literature record of this species. The white dot indicates the locality where a specimen of this species was collected. From the comparison it results that museum data expand the known geographical distribution of the species.*

In the Ecology web page of each registered species, a possible user can acquire the most important literature references concerning the species habitat, its autoecology and synecology. Furthermore, a dynamic graph is created giving the vertical distribution of the species as it results from the literature information, stored in the database, and the museum collection data (Fig. 4). From the evident comparison a user can assess whether the museum data expand the known vertical distribution of this species.

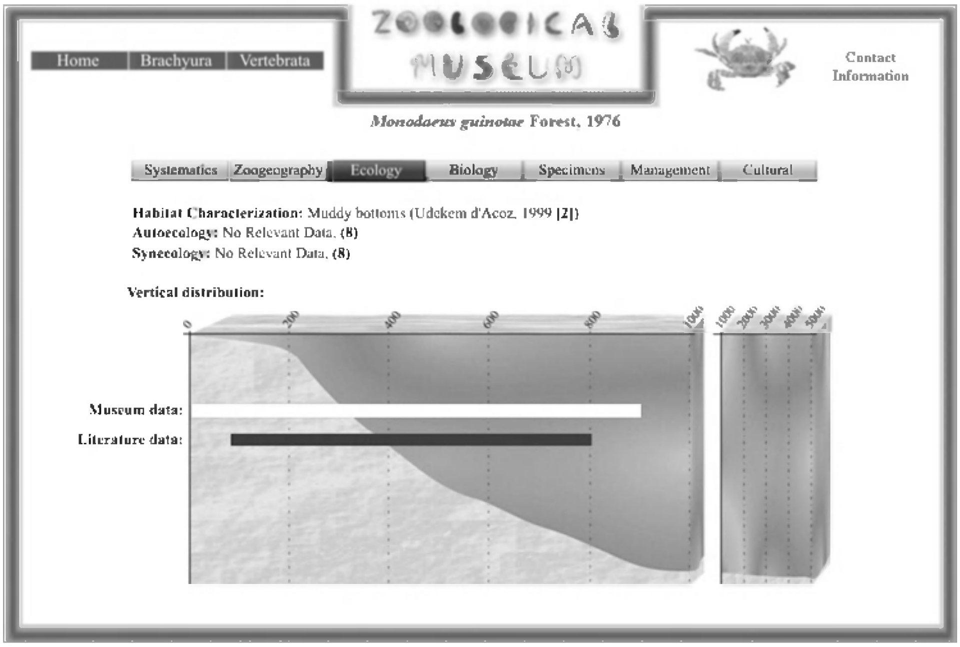


Fig. 4. The "Ecology" web page for the species Monodaeus guinotae Forest, 1976. Dynamic chart showing the vertical distribution of this species according to literature data (black bar) and the museum collections data (white bar). Comparing these two bars it results that the museum data expand the known vertical distribution of the species.

In the Biology web page, references on the life cycle of the certain species, its size, its feeding habits, possible predators and diseases are given, while as in all thematic web pages, links to other relevant internet resources are provided.

A table providing all the available specimens of the species selected, which can be found in ZMAUTH collections, mainly structures the Specimens web page. For each museum specimen, the museum code number, information on the station where it was collected as well as the number of individuals contained, are given. From this web page a possible user can ask for a loan or for more information about a specimen by clicking a relative link which directly opens the e-mail client.

Following the Management link from a decapod species main page a new web page is created containing the most important literature information concerning the harvest of the certain species, its cultivation and production as well as other possible uses.

One of the most important innovations of this information system is the provision of existing cultural information about each registered species. In the respective web page, a possible user can find the available historical references of this species (*e.g.*, possible reference from Aristotle), mythological references as well as the contribution of this species in literature and fine arts. Finally, references for possible threats for a certain species by human activities are also provided.

Utilization of the information system in research and education

This information system can prove to be a very useful scientific and educational tool due to its structure, the contained information and its updating possibilities as well.

An example of the use of this information system is the acquisition of new information on the vertical distribution of a decapod species or a group of species. When a new specimen of a certain species is registered in the database, the museum data bar in the corresponding dynamic chart in the species ecology web page (Fig. 4) is updated according to the sampling station's depth. Through the comparison of this bar with that resulting from the relevant literature information, it can be assessed whether the known vertical distribution of the certain species is expanded in shallower or greater depths. Thus, through this procedure new scientific information is acquired.

A second example concerns the potential acquisition of new information on the geographical distribution of decapod species which are Lessepsian migrants. When a specimen of Lessepsian migrant species is collected and registered in the database, the position of its sampling location will appear in the corresponding dynamic map of the Eastern Mediterranean which also depicts all previous locations where this species has been recorded till now (Fig. 3). By the evident comparison of the geographic location of the sampling station with the previous known distribution of this Lessepsian species, a possible user can assess whether the known geographical distribution of this species is expanded and hence if new information is acquired.

This information system can also be a very useful medium for educational purposes. It can be integrated in courses dealing with zoology, marine biology or zoogeography. Through the static web pages students can be trained in the basic concepts of invertebrate systematics, while the resources contained in the dynamic web pages can contribute to a better training and understanding of the ecology and zoogeography of the decapod fauna of the Mediterranean and the NE Atlantic.

Future plans

The Zoological Museum of the Aristoteleio University of Thessaloniki hosts large collections of many other invertebrate groups (*e.g.*, Porifera, Actiniaria, Polychaeta, Peracarida) which comprise more than 2,000 species. In the future, all these species will be registered in this information system along with the relevant literature information. All these data will be available through the web site of the Zoological Museum (A.U.Th) using the same presentation scheme as it has been done for the decapod species. This information system intends to be a single source of information on the marine invertebrate diversity of the Aegean Sea and to contribute to the free dissemination of the relative information in the international scientific community and to the development of educational tools for a more comprehensive training of the students in the fundamental concepts of biodiversity, zoogeography and ecology of marine invertebrates.

References

O'Connell A.F. Jr., A.T. Gilbert and J.S. Hatfield. 2004. Contribution of natural history collection data to biodiversity assessment in national parks. Conserv. Biol. 18:1254-1261.

Ponder W.F., G.A. Carter, P. Flemons and R.R. Chapman. 2001. Evaluation of museum collection data for use in biodiversity assessment. Conserv. Biol. 15:648-657.

Suarez A.V. and N.D. Tsutsui. 2004. The value of museum collections for Research and Society. Bioscience 54:66-74.

Udekem d'Acoz C. 1999. Inventaire et distribution des Crustacés Décapodes de l'Atlantique nord-oriental, de la Méditerranée et des eaux continentales adjacentes au nord de 25°N. Patrimoines naturels. Museum national d'Histoire naturelle, Paris, 383p.

Dealing with the challenges of presenting taxonomic data online: An introduction to PLANKTON*NET@AWI

Alexandra Kraberg, Friedrich Buchholz and Karen H. Wiltshire

Alfred Wegener Institute for Polar and Marine Research, Biological Station Helgoland, Kurpromenade 201, 27498 Helgoland
E-mail: akraberg@awi-bremerhaven.de

Abstract

Phytoplankton taxonomy requires comparison of large volumes of information including images of taxa from different geographical areas. The internet should be ideally suited for this task. However, despite its advantages compared with traditional dissemination methods and the huge array of different online taxonomic resources, it lacks the evaluation and validation mechanisms of traditional resources, and 'ground rules' for the treatment of taxonomic data have not yet been established. PLANKTON*NET@AWI contains more than a thousand plankton images from the North Sea and different collections from all over the world. The database can be searched alphabetically or via collections. Each record can be viewed as a standardized data sheet with images and taxonomic descriptions. Comment functions are also provided but their administration has yet to be discussed. Images from different collections can be compared, facilitating the detection of taxonomic inconsistencies and geographic variations in morphology. PLANKTON*NET is a collaborative project with partners at Roscoff and in Woods Hole, but the individual sites are not yet networked. We are currently exploring mechanisms for future database formats and ways of networking existing resources to maximize the benefits for taxonomic research. Our favoured approach will be to follow the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) as an application-independent interoperability framework based on XML technology, so that the integrity of local taxonomic initiatives can be maintained, while sharing content but this will also require discussion within the wider scientific community.

Keywords: Taxonomic database; long-term data; data analysis; OAI.

Introduction

Global change phenomena have already been shown to affect biological communities. Conclusions as to what exactly the long-term consequences of global change events on communities in different types of ecosystems might be vary widely, but there is some agreement that these impacts will be profound. One reason for the difficulties in analyzing and predicting long-term trends is that such analyses critically depend on reliable biodiversity data about the ecosystem under study. However, such baseline taxonomic information is often lacking and how environmental changes will affect species, communities and the stability of the affected ecosystems in general is therefore difficult to predict (Danielsen, 1997; Fjeldså and Lovett, 1997; Gray, 2001; Piraino *et al.*, 2002).

Filling gaps in the knowledge about the number of species within a given geographic area would normally be the domain of trained, specialized taxonomists. However, there is today only a relatively small number of such specialists, facing millions of unknown taxa particularly in marine systems. In addition, the traditional publication process is very complex and expensive excluding many potentially interested parties from quick access to new information that could be vital for the design of new scientific programmes (Agosti and Johnson, 2002; Godfray, 2002).

Trained taxonomists are becoming rare for many species groups and many taxonomic collections that have been built over years or even decades are not maintained beyond the retirement of the scientists who assembled them. This together with slow and patchy access to existing and emerging information will lead to the loss of vital information and duplication of research efforts. This is happening at a time where financial resources are often stretched and where concurrently there is a renewed need for taxonomic expertise as entirely new habitats *e.g.* in the deep-sea are discovered (Morin and Fox, 2004). Online resources, if properly organized and co-ordinated could be one way out of that dilemma (Costello, 2003). This paper presents PLANKTON*NET@AWI a resource that is developing a concept for not only site design but also participation of the scientific community to provide a widely accepted, easy to implement and cost effective format that can deal with the problems described above. Once fully implemented PLANKTON*NET@AWI can complement and enhance traditional data dissemination methods.

Methodology and Data Acquisition

The database set-up is based on mySQL/ PHP which is open source software and easy to install and run. Further modules are based on XML and we are trying to establish a pilot for links using open archives initiative meta harvesting protocols (OAI-MHP) with our existing partners at the Station Biologique de Roscoff, who are hosting a second PLANKTON*NET site with their own data.

Data acquisition for PLANTKON*NET is essentially a collaborative process and the PLANKTON*NET developers and administrators are aiming at developing a resource that can act as a custodian for the data of different data providers, where on behalf of the data provider, PLANKTON*NET is handling all data processing and their addition to the database as a collection of images and taxonomic descriptions. In return these data providers will have a say in the further development of this resource. Individual contributors' records are clearly identifiable in the database and each contributor can provide additional information for the design of a customized introduction page that describes the institute at which the data were collected, provides geographical information about the sampling origin etc. The contributor wherever possible, provides image information in .tif format. These are then processed and will be used as jpg throughout the site and as .tif for download versions of all of the images. Where necessary, PLANKTON*NET staff will also process original image material such as slides and photographs for the contributors, whenever they lack the resources to do so.

PLANKTON*NET aims and set-up

A major aim of PLANKTON*NET is to network biodiversity resources but at the same time to allow individual sites the freedom to develop modules and tools according to their own needs. In the case of the Biologische Anstalt Helgoland for instance (which is part of the Alfred-Wegener-Institute for Polar and Marine research), there is a strong emphasis on researcher and student training and this will be reflected in the analysis tools already planned for PLANKTON*NET@AWI.

At the heart of PLANKTON*NET lies its classification system, linked to the Taxonomic name server (TNS), originally developed at uBio, Woods Hole. The underlying databases of the TNS indexing system, contains approximately 1.7 million taxon names, synonyms and vernacular names in addition to valid scientific names, allowing quick and even more importantly complete assemblage of the taxonomic and other literature about a given taxon.

The taxon information within PLANTKON*NET is organized around collections (Fig. 1 and 2) each of which has been provided by one collaborator and is easily identifiable throughout the PLANKTON*NET site as originating from that collaborator. From a list with all collections in the database, a list with the genera present in the individual collections can be accessed.

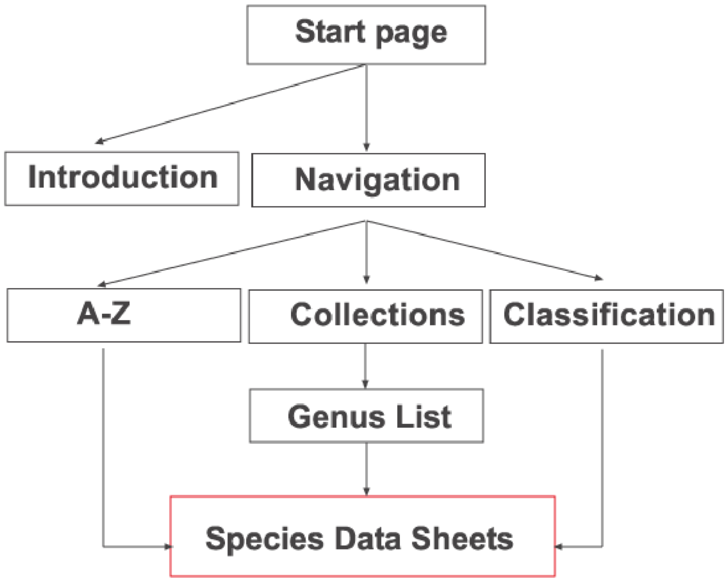


*Fig. 1. The PLANKTON*NET structure: Site Map with components and navigation of the PLANKTON*NET@AWI taxonomic web resource.*

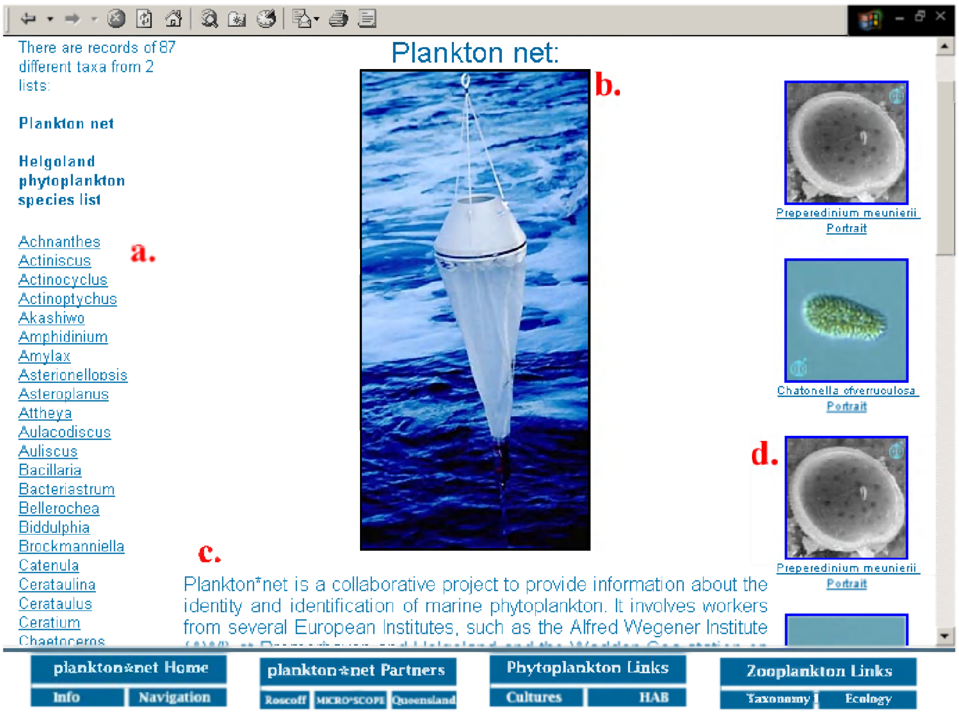


Fig. 2. A start page for an individual collection (accessed from the collections list): a list of all genera present in the collection; b. image or logo chosen by the data provider; c. introductory text for the providing institute and the image collection; d. examples of images contained in the collection.

From these lists the user can navigate to individual species sheets (Fig. 3), containing enlarged images, information about image origin and short species descriptions.

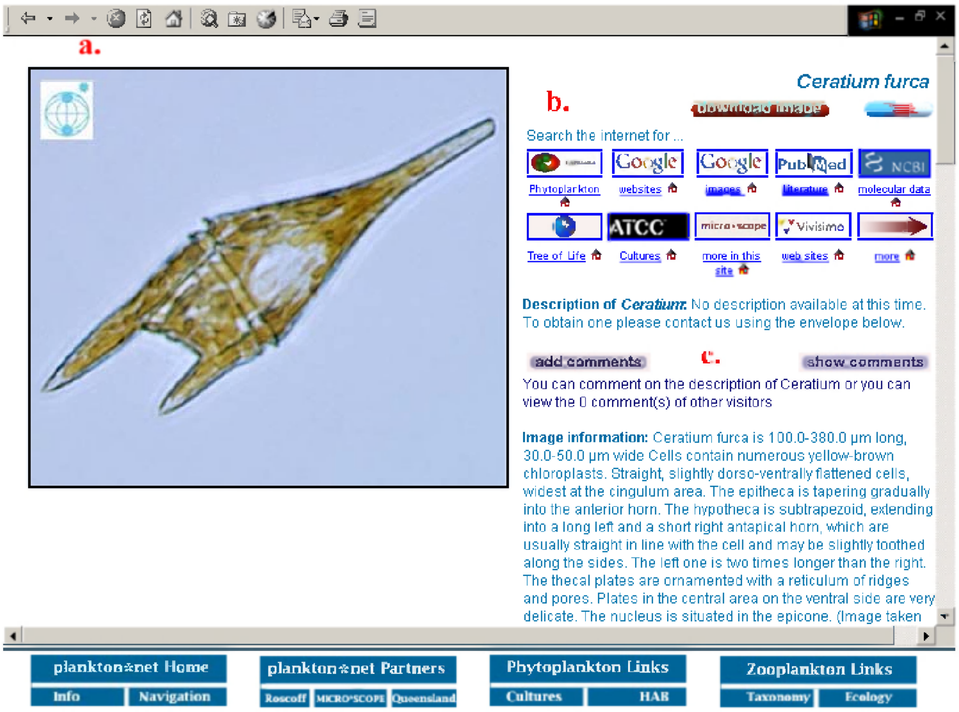


Fig. 3. Example of the individual species pages, as accessed from A-Z index and collections pages. The main components are: a. Large image of the organism; b. sets of outlinks with deep links to external resources containing information about the organism; c. image information and species descriptions.

These species sheets also contain a series of outlinks that connect to external online resources with information about the organism in question. These are deep links, *i.e.* they do not connect to a homepage, that has then to be navigated, but directly to the relevant information within these external websites. The outlinks have been customized so that, for instance species pages about Harmful algal bloom (HAB) species contain outlinks to HAB relevant rather than general phytoplankton resources. In addition, each species page contains a link to a high resolution download version of the image (in .tif format). These can be downloaded free of charge unless they are intended for commercial use. In addition, the bottom section of the species sheet contains a collation of all images of the same species and genus within PLANKTON@NET (Fig. 4), this simple set-up already allows limited comparisons of geographical images and importantly also facilitates the detection of possible species identification errors.

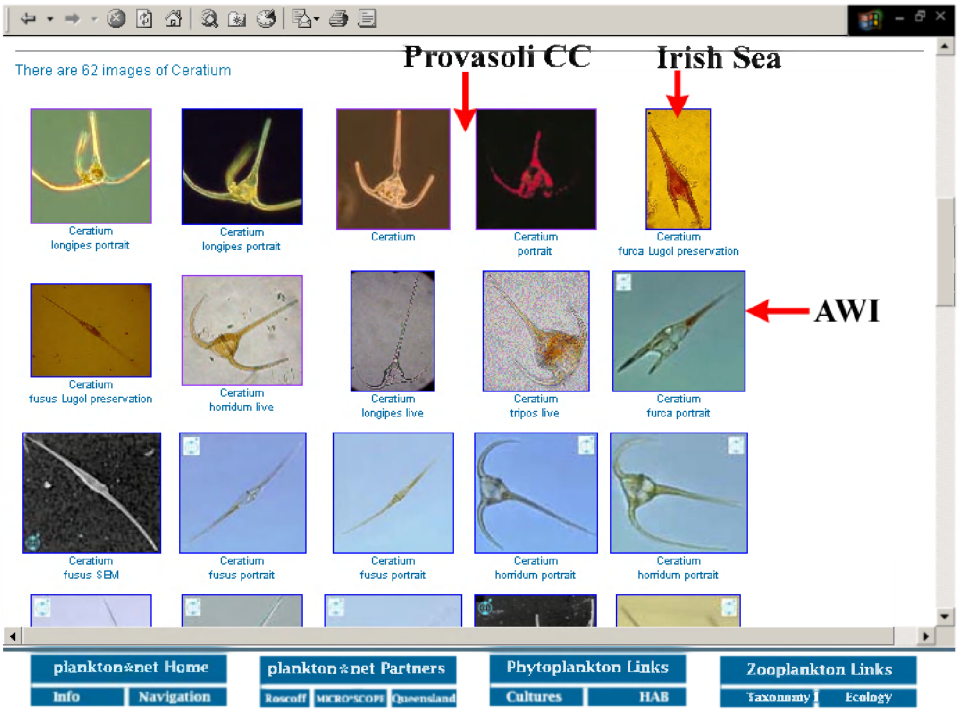


Fig. 4. The lower part of the individual species sheet, showing all images for this species present in the database. The example given here includes images from 3 different geographical sources.

While the above features are served out of PHP scripts/ MySQL database queries, there are also a number of static features in PLANKTON*NET, for instance a link list with links to culture collections as well as further online databases, taxonomic resources and identification aids. These static resources can of course be customized to the requirements of individual PLANKTON*NET sites.

Discussion and outlook

At present the emphasis of PLANKTON*NET is on the display and the collation of data. However we think that it will be vital to extend this basic format to be more conducive to data analysis. The challenge now is therefore to increase functionality so that the resource can be used by as wide a range of user groups as possible while maintaining and enhancing its user friendliness without disrupting site organization. Three stages are planned in the further development of PLANKTON*NET. First of all the current site structure will be modified to increase search functionality at species level. Although at present individual data sheets appear with a species heading, the site still essentially operates at genus level, which becomes apparent in Fig. 4. Such an increase in functionality could mean splitting the taxonomic descriptions into several separate fields to allow more refined searches. A key development will be to work on achieving interoperability between PLANKTON*NET nodes and with already existing resources. The third step will be the addition of modules such as a library of taxonomic keys. These will be xml based matrix keys. By applying the emerging TDWG standard (Taxonomic Databases Working Group) for representation of descriptive data, (structure of descriptive data, SDD), PLANKTON*NET@AWI will be able to ensure that information is usable with different key production software tools, again allowing flexibility for local initiatives.

An important point to stress is that this resource is not in competition to existing resources. On the contrary, it is vital to maintain dialogue with these initiatives to maintain interoperability between resources. But with its clear focus on plankton taxonomy and in addition the analysis and interpretation of plankton data, PLANKTON*NET@AWI can provide a future essential tool for phyto- and zooplankton taxonomists world wide and can be used as a standalone tool or in conjunction with other existing resources, so that all the information necessary for gathering reliable biodiversity information, can be easily assembled and used.

The main underlying principle is to produce a resource that can be used on a par with traditional paper publications. This requires several prerequisites. Firstly the site structure has to be such that it is transparent, *i.e.* with all information necessary to judge the quality of the information presented. Secondly the format has to be supported and recognized as a valid format by the scientific community. It is for this reason that the PLANKTON*NET project is placing considerable emphasis on collaboration and communication with external experts and data providers to ensure that this resource will be of the utmost relevance to scientists in general, not only taxonomists and that importantly it is maintained in the long term (despite the financial pressures placed on most of the participating institutes). It is therefore vital that PLANKTON*NET is inclusive at every point of its development and is seen to encourage participation of additional data providers in the future. Providing a clearly and logically structured coherent site structure will aid this process.

The PLANKTON*NET concept allows scientists holding taxonomic collections, to have them preserved with minimum investment, while still having a say in the treatment of their taxonomic material within the database. With an increasing number of contributors this will become not only a means of preserving taxonomic information and the integrity of individual taxonomic collections but also, and importantly, a means of quality control of the data sets within PLANKTON*NET. This will allow the production of a reliable resource without the need for restrictive and, above all, time and money consuming editorial structures, but yet trusted in the same way that traditional paper based publications are. This resource is in no way meant to replace conventional resources but to complement them. The challenge of this and other future resources will be to achieve this without restricting one of the greatest strengths of the worldwide web: the ease and speed of the cost effective transfer of scientific information worldwide.

Acknowledgements

The authors, on behalf of the PLANKTON*NET community, wish to thank the developers of the original software, first at the University of Sidney and now at the Marine Biological Laboratory at Woods Hole. The actual data providers also deserve our sincerest thanks for the taxonomic information provided as well as for their comments and suggestions.

References

Agosti D. and N.F. Johnson. 2002. Taxonomists need better access to published data. Nature 417:222.

Costello M.J. 2003. Role of on-line species information systems in taxonomy and biodiversity. In: Heip C.H.R. (ed) Biodiversity of coastal marine ecosystems patterns and processes: a Euroconference, Corinth, Greece, 05-10 May, 2001, p58-59.

Danielsen F. 1997. Stable environments and fragile communities: does history determine the resilience of avian rainforest communities to habitat degradation? Biodiversity and Conservation 6:421-431.

Fjeldså J., J.C. Lovett. 1997. Biodiversity and environmental stability. Biodiversity and Conservation 6:315-323.

Godfray C.H.J. 2002. Challenges for taxonomy. Nature 417:17-19.

Gray J.S. 2001. Marine diversity: the paradigms in patterns of species richness examined. Scientia Marina 65:41-56.

Morin P.J., J.W. Fox. 2004. Diversity in the deep blue sea. Nature 429:813-814.

Piraino S., G. Fanelli, F. Boero. 2002. Variability of species'roles in marine communities: change of paradigms for conservation priorities. Marine Biology 140:1067-1074.

Dataset and database biodiversity of plankton community in Lebanese seawater (Levantine Basin, East Mediterranean)

Sami Lakkis

Section of Oceanography, Lebanese University, Beirut, Lebanon
Marine Bio-Consultant , P.O.Box 138, Byblos, Lebanon
E-mail: slakkis@ul.edu.lb

Abstract

Oceanographic data were obtained from ship cruises conducted from 1965 until 2003 in the neritic and oceanic Lebanese waters (Levantine Basin). They include, in addition to Plankton community diversity and abundance, main hydrographic data such as temperature, salinity, dissolved oxygen, water transparency, chlorophyll a, nitrates and phosphates. The purpose of this work was to elaborate a dataset and database for the biodiversity of the plankton community in relation to the hydrological conditions of the area and the abundance of species populations. Spatial and temporal qualitative and quantitative distributions of the species and groups are strongly correlated to seasonal variations of hydrological parameters. The thermal annual cycle splits up in two different phases: a cold phase in winter (December-March) and a warm phase in summer (June-November), separated by a short spring inter-season. During the winter period, isothermal conditions prevailing in the 0-200 m water column are characterized with relative low temperature and salinity averages, poor plankton productivity and low biomass contrasting the high species diversity. The warm phase in summer is characterized by high surface water temperatures and salinity, accompanied with the formation of a thermocline in the layer 35-75m and a water layer stratification, which creates a hydro-thermal barrier avoiding exchange of water masses and vertical migration of organisms and low plankton biomass and diversity. A short spring period (April-June) is characterized by optimal hydrographic conditions that induce phytoplankton growth leading sometimes to little blooms, followed by high zooplankton production contrasting the low species diversity. During the long-term survey we noticed certain hydrological changes in the Levantine Basin, expressed by small rises in temperature (ΔT=0.4°C) and salinity (ΔS= 0.35‰) and increasing migration. To date 400 phytoplankton species and more than 750 zooplankton taxa, were identified in the area, of which dozens of introduced exotic species are of Indo-Pacific origin. These changes induced some ecological evolutions in the marine ecosystems and are due to the regulation of the water level of the Nile by the Aswan High Dam by reducing the amount of freshwater outflow in the Mediterranean, and the deepening of the Suez Canal accelerates the northward current and thus facilitates the migration process. Global warming may also contribute to these hydrological changes in generating a certain "tropicalisation" of the Levantine Sea.

Keywords: Lebanese waters; Levantine Basin; Plankton datase; Biodiversity.

Introduction

The Eastern Mediterranean, particularly the Levantine Basin, is the most impoverished and oligotrophic water body in terms of productivity and nutrient concentration (Krom *et al.,* 1991). There is a well-defined eastward trend in nutrient ratios over the entire Mediterranean that starts at the Gibraltar straits, and continues through the western basin, towards the eastern basin. The supply of nutrients to the Mediterranean is limited by inputs from the Atlantic Ocean and those of various rivers surrounding the sea (Hecht and Gertman, 2001). These authors have found that the surface layer to a depth of 150m had a nitrate concentration of less than 1µmole.l^{-1} and the maximum reached was 5.4 - 6.5µmole.l^{-1}. Comparing the nutrient distribution and budget in the Mediterranean and the Red Sea, Souvermezoglou (1988) suggested that the Mediterranean receives 70% of its nutrient supply from the Atlantic, the rest of nutrients being provided by the rivers. The low concentration of nutrients, namely phosphates, nitrates and silicates induces low primary production due to poor chlorophyll content (Berman *et al.*, 1984). The strong evaporation and the shortage of freshwater input, makes of the Levantine Sea a concentration basin with a relative warm and high saline water body, where the temperature and salinity are the highest of the whole Mediterranean (Lacombe and Tchernia, 1972). The water masses mixing during homothermic winter period provide nutrients from the deep layers to the euphotic zone (Dugdale & Wilkerson, 1988; Ediger *et al.*, 1999). It has been demonstrated that the near-bottom chlorophyll-maxima are recorded in summer between 90 and 120m below the thermocline and above the sediments and over the continental shelf in the southeast Mediterranean off the Israeli coast (Townsend *et al.,* 1988); The bottom sediments can constitute the main source of nutrients for picoplankton and nanoplankton development and thus for chlorophyll-a and zooplankton production.

Very little was known about the Levantine Intermediate Water (LIW) and deep circulation in the Eastern Mediterranean until 1991. After the POEM results were revised and added to the Meteor expedition 1995, new findings were introduced regarding a big transition in the circulation of LIW. Malanotte-Rizzoli *et al.* (1998) have suggested that the LIW formed inside or in the periphery of Rhodes gyre is blocked in its westbound route by a three-lobe strong anticyclonical structure in the Southern Levantine, which induces a substantial LIW recirculation in the Levantine Basin itself.

The general circulation along the coast of Lebanon is dominant in northward direction during most of the year, in accordance with the general counterclockwise current gyre of the Eastern Mediterranean. This current is locally modified by the configuration of the coastline and the topography of the narrow continental shelf. This results in a series of clockwise directed eddies and small gyres associated with bays and headlands as well as with numerous submarine canyons incised in the continental shelf (Goedicke, 1972). Water movements along the coast are strongly associated with surface currents and seasonal meteorological factors. More detailed hydrographic data and its relation to plankton biodiversity were reported in previous work (Lakkis *et al.,*1996; Lakkis, 1997).

The spatial and temporal variability in plankton communities is mainly correlated to seasonal hydro-climatic factors prevailing in the area. Seasonal changes in the quality and quantity of plankton are very pronounced, whereas inter-annual fluctuations are more regular and show little variability (Lakkis and Novel-Lakkis, 1981).

The purpose of this paper is to highlight a plankton dataset and biodiversity in the Lebanese sector over a long-term survey 1965-2003, along with seasonal variations and inter-annual fluctuations.

Methods and materials

The study area extends over 150 km between the south of Lebanon (33°42'-34°28'N) and offshore Tripoli city in the north (35°27'-35°31'E), covering a total of 46 stations in the neritic waters and offshore in deep oceanic area (Fig. 1). Monthly, seasonally and annual cruises were carried out between 1965 and 2003. Surface and vertical plankton samples accompanied with hydrology measurements were taken, including plankton nets tows, temperature, salinity, dissolved oxygen, phosphate, nitrate, chlorophyll-***a***, and pH. Samples were taken at nine standard depths 0m, 10m, 25m, 35m, 50m, 75m, 100m, 150m, 200m.

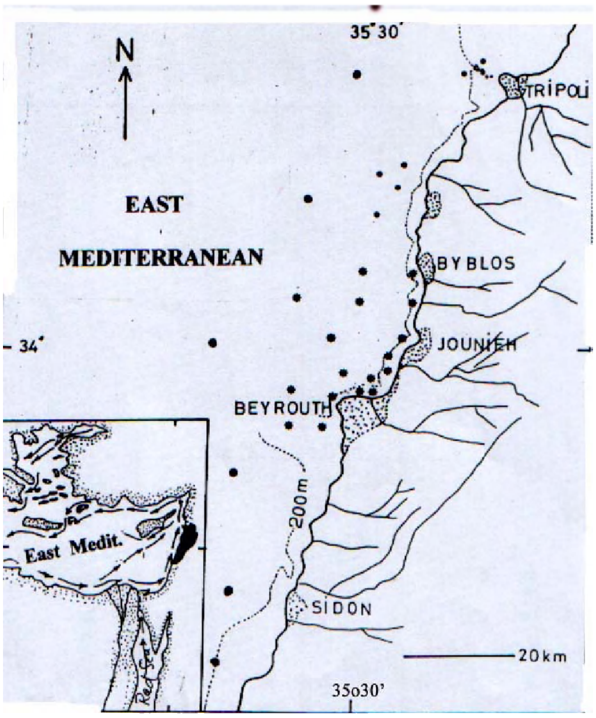


Fig. 1. *Location of sampling stations along the coast of Lebanon during 1965-2003. Large spots: monthly and seasonally visited stations; small spots: irregularly or occasionally stations. Contour of the narrow continental shelf is indicated by dotted line. The insert indicates the general circulation in the Levantine Basin.*

Monthly and seasonally cruises carried out during 38 successive years have provided a total of 2399 zooplankton samples and 1200 water samples for phytoplankton, chlorophyll-a and nutrient analysis (Figs. 2, 3). STD Hydro-bios electronic probe was used for measuring in situ temperature and salinity; and an Oxygen-meter and pH-meter probes were used for measuring dissolved oxygen and pH data. Niskin and Nansen reversing bottles were used to collect water samples for chemical analysis and CHL-***a*** detection/analysis. Nitrate and phosphate were analyzed according to Strickland and Parsons (1972); while CHL-***a*** was determined using Parsons (1969) for the determination of photosynthetic pigments in seawater. Vertical tows of coupled plankton nets of 50μ (for microplankton) and 200μ (macrozooplankton) were subject to qualitative and quantitative analysis.

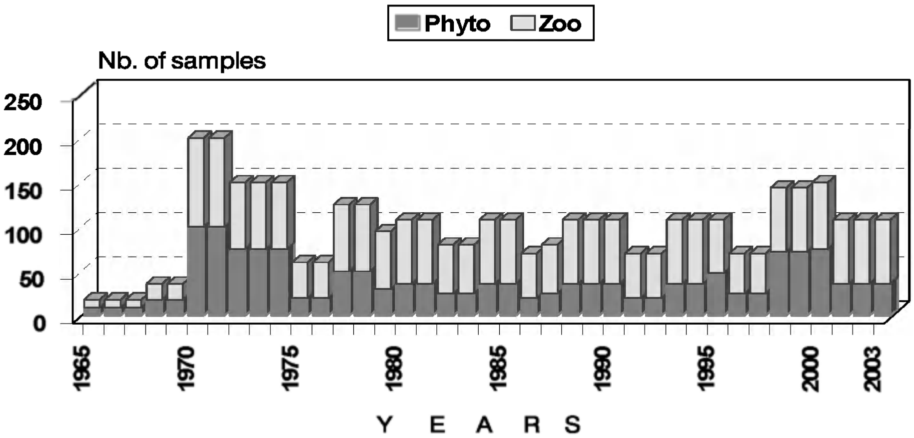


Fig. 2. Number of zooplankton samples (light column) and phytoplankton (dark) collected in the Lebanese seawaters during between1965 and 2003.

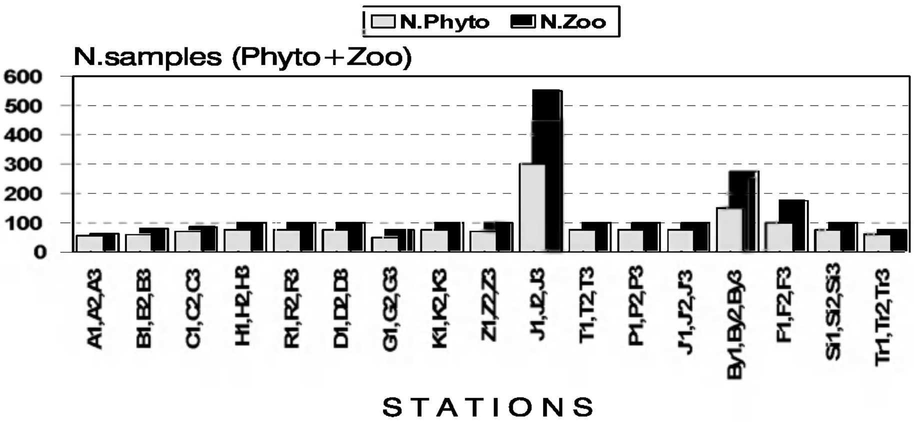


Fig. 3. Total Number of zooplankton samples (dark column) and phytoplankton (white) per station during1965-2003.

Taxonomic identification was done up to species level; the abundance was reported as cells.l^{-1} for phytoplankton and chlorophyll a in mg.m^{-1}, whereas the zooplankton was reported as number of organisms per m^{-3} and/or cc.m^{-3} and mg.m^{-3} of dry weight of wet weight. Details of oceanographic data were reported in the IOC/EU project of MEDAR/MEDATLAS II (Maillard *et al.*, 2001; Lakkis, 2002).

Dataset

The plankton samples and hydrographic profiles taken during the long-term survey (1965-2003) constitute the basic elements for the Biodiversity database of Plankton community.

Results

Hydrological properties of the Lebanese seawater

Two annual thermal phases characterize the Levantine basin including Lebanese waters: a cold phase in winter (December-March) and a warm phase during the hot and dry long summer (June-November). A short spring inter-season separates the two periods.

Cold phase

The cold phase corresponds with the winter season (December-March); it is characterized by relative cold seawater (17°C) with isothermic conditions in the water column, due to mixing and turnover of water masses. Big amount of freshwater input from runoff and rivers reduce the surface salinity at offshore stations to its minimum in February-March (38.95±0.41) (Fig.4).

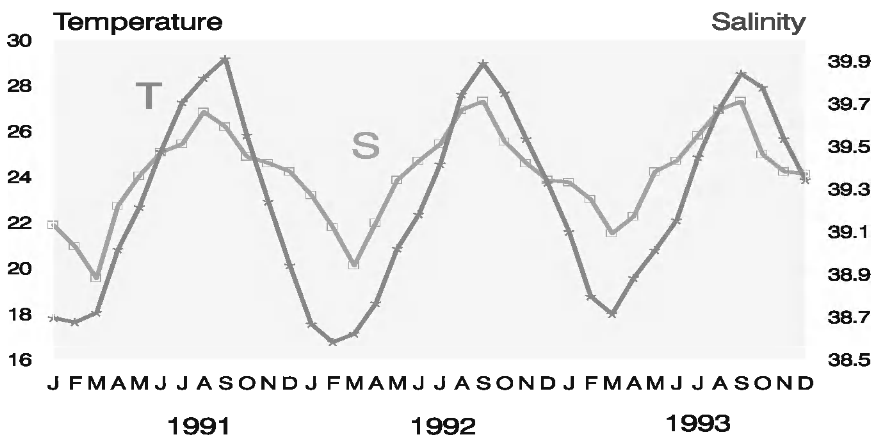


Fig. 4. Seasonal variations of Temperature and Salinity at surface offshore Lebanese seawater during three successive years at one offshore station J2.

Concentrations of nutrients reach their maximum levels in January ($PO4=0.25$ µmole.l^{-1}; $NO3=0.33$ µmole.l^{-1}); whereas CHL-***a*** is low during this period, averaging 0.09±0.04 mg.m^{-3} for the season. Phytoplankton standing crop is very poor and zooplankton biomass is at lowest annual values, contrasting with high species diversity (Lakkis, 1997) .

Warm phase

This phase coincides (if past tense, there should be past tense in cold phase and probably a year) with the hot long period of summer (June-November), during which the thermocline is formed between 35 and 75 m, accompanied by water column stratification. Surface water temperature increases to reach its maximum in August or September with a peak of 30°C. Salinity reaches a maximum of 39.62‰ ± 0.15 in September at offshore stations, which is the highest value in the entire Mediterranean. Low phosphate and nitrate concentrations during this period, coincides with the lowest CHL-***a*** of the whole year (0.07±mg.m^{-3}±0.05). Dissolved oxygen at the surface drops to its lowest level in August (5.72 ml.l^{-1} ±0.03) due to rising temperature.

Spring Inter-season

This short phase corresponds to the spring season (April-June) and is characterized by optimal hydrological conditions suitable for phytoplankton growth, leading sometimes to a little bloom at coastal water with high CHL-***a*** values (0.41mg.m^{-3}±0.12). Water temperature varies between 21°C in April and 23°C in June, while moderate salinity varies between 39.35‰ and 39.42‰ respectively. Dissolved oxygen in the euphotic zone is correlated to high phytoplankton standing crop. Nutrient concentrations drop to minimum averages in June for phosphate (0.05µmole±0.01) and in May for nitrate (0.17µmole±0.09). This decrease is due to the use of nutrients by the microalgae for their growth.

Taxonomic Diversity of Plankton Community

Plankton community of the central Levantine Basin, although impoverished, remains diversified; most of Mediterranean groups of species are present. Several species of Indo-Pacific origin have migrated through the Suez Canal to establish permanent populations confined to the Eastern Mediterranean. Few of them have in mean time extended to the Western Basin.

Phytoplankton

About 400 phytoplankton species were found in the Lebanese waters, including 160 diatoms and 230 dinoflagellates. Some silicoflagellates were also identified**.** The composition of the phytoplankton populations varied in space and time. Several species occur permanently; they are found all year round, whereas many others are encountered only during some months of the year (Lakkis & Novel-Lakkis, 1981). The major taxa are given in Fig. 5 and Table III.

Table III: Taxonomical composition of Phytoplankton community.

Group	Families	Genera	Species
Bacillariophyceae	15	46	151
Dinoflagellata	14	33	227
Silicoflagellata	3	3	5
Ebriidae	2	2	2

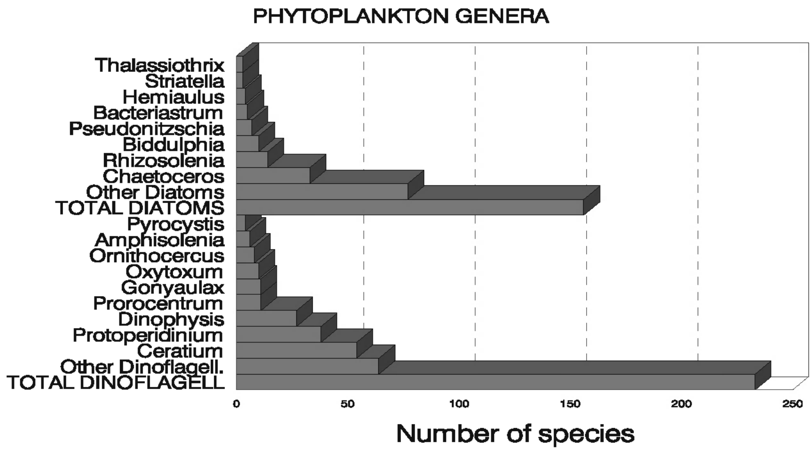


Fig. 5. Taxonomic Diversity of Phytoplankton community in Lebanese seawaters during 1965-2003.

The seasonal distribution of the most common and more frequent species is as follows:

Winter: Chaetoceros curvisetrus, Ch. pseudocurvisetus, Ch. decipiens, Leptocylindrus danicus , Skeletonema costatum, Pseudonitzschia fraudulenta, P.seriata,Cerataulina pelagica, Dinophysis caudata, Protoperidinium divergens, P.diabolus.

Spring : Ch. pseudo-curvisetus, Skeletonema costatum, Leptocylindrus danicus, L. minimus, P. fraudulenta, P. seriata, P. pungens, P. closterium

Summer-Fall: Chaetoceros affinis, Ch. brevis, Ch. didymus, Ch. Anastomosans, Ch. rostratus, Streptotheca thamesis, Rhizosolenia calcar-avis, Bacteriastrum elegans, Ceratium furca, C. pulchellum, Dinophysis caudata, Protoperidinium divergens, P. Diabolus, Dinophysis caudata, Prorocentrum micans.

Monthly changes in species diversity and abundance of phytoplankton depend on hydro-climatic factors and hydrological conditions. At offshore stations, surface temperature ranges between 16°C in February, and a maximum of 30°C in August. Seasonal variability of salinity is small; it ranges between 39.29‰±0.49 in March and 39.62‰±0.34 during September). Nutrient concentrations display strong monthly variations at the surface, with a maximum in January for phosphate (0.25µmole±0.05) and nitrate (0.33µmole±0.15) and a minimum in June (phosphate) and May (nitrate) CHL-**a** displays great seasonal variations, the minimum average is recorded in September (0.07mg.m^{-3}) and maximum in May (0.39mg.m^{-3}) at open sea.

The annual cycle of phytoplankton displays two peaks: a major one in May-June and a 2nd less important one in October-November. In summer, during the water stratification, the plankton at the surface water is very poor in quantity as well as in diversity (Fig. 6). At 10-25 m the phytoplankton is more abundant and more diversified than at the surface. In winter the isothermic conditions and the turnover of water masses are not suitable for phytoplankton growth, which keeps densities at a low level.

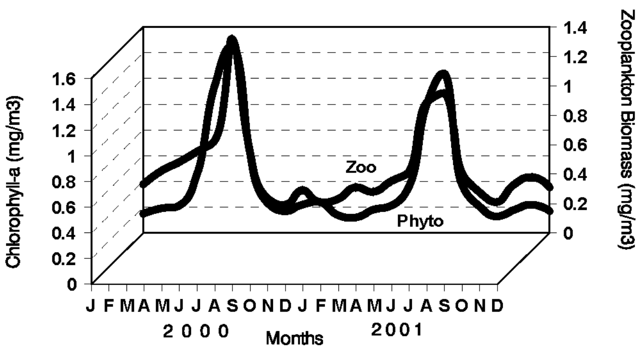


Fig. 6. Seasonal variations of chlorophyll a and Zooplankton at surface offshore station J2 during 2000-2001.

During spring (April-May) phytoplankton reach their maximum abundance ranging between 50.000 and 100.000 cells.l^{-1}. At the same time, the diversity is very low, because of the dominance of few species. In summer, following the breakdown of spring phytoplankton bloom, the cells density drops to 3000-10.000 cells.l^{-1}. Abundance of phytoplankton and chlorophyll-a shows a decreasing gradient from coastal waters towards offshore deep waters. Regarding the vertical distribution, the phytoplankton is abundant at the upper 50 meters and decreases drastically to disappear at 100m. However during summer and fall, the maximum depth for chlorophyll is recorded over the bottom sediment at 90-120 m due to source of nutrients from the regeneration of benthic organic substances.

Zooplankton

To date about 1000 zooplankton taxa were identified in the Lebanese seawaters, including all the planktonic groups from protozoans up to prochordates (Larvacea and Fish larvae). 250 microplankton species were found, of which 141 Tintinnids, 25

Foraminifera, 10 Acantharia, 25 Radiolaria Spumellaria, 30 Nasselaria, 6 Phaeodaria and 1 Heliozoa; many of them are of Indo-Pacific origin, present in the Red Sea.

Macrozooplankton includes all species from Hydromedusae up to Tunicates and fish larvae (Fig. 8, Table 4). The zooplankton community shows pronounced seasonal variations in diversity and abundance. A close phytoplankton-zooplankton relationship is always observed in the area (Fig. 7). The seasonal distribution of zooplankton is resumed as follows:

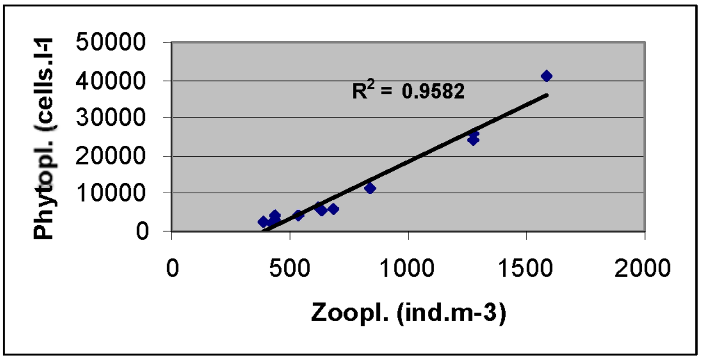


Fig. 7. Phytoplankton-Zooplankton relationship in the Lebanese waters during 2000 at offshore station J2.

Table 4. Taxonomical composition of Zoolankton community in the Lebanese seawaters during 1965-2003.

Group	N. species	Group	N. species
Foraminifera	12	Mysidaceae	4
Actinopoda	66	Cirripedia (larvae)	4
Tintiniidae	141	Decapoda (larvae)	110
Hydromedusae	68	Chaetognatha	10
Scyphozoa	5	Pteropoda	9
Siphonophora	28	Heteropoda	4
Copepoda	173	Polycheata (larvae)	8
Cladocera	6	Polycheata (adults)	4
Ostracoda	6	Appendicularia	15
Amphiopda	25	Thaliacea	6
Euphausiacea	5	Eggs & fish larvae	90

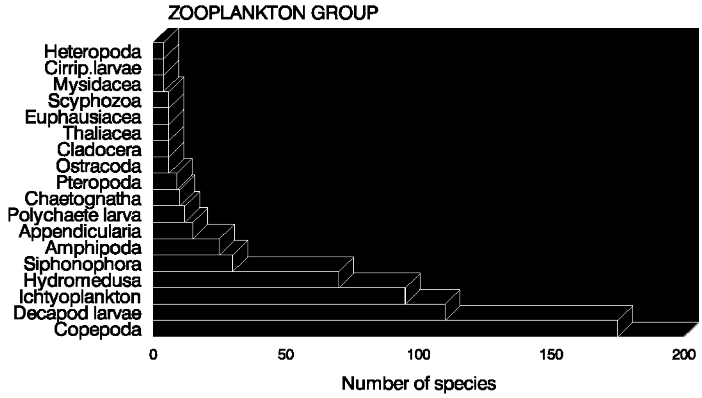


Fig. 8. Taxonomic Diversity of Zooplankton community in Lebanese seawaters during 1965-2003.

Winter: During the isothermic conditions, the zooplankton occurs inlow abundances, but with high species diversity. Most of the species found at the surface are mesopelagic forms and are carnivorous, for example large Copepods, Siphonophora, Hydromedusae, Chaetognatha, etc…

Spring: With optimal temperature, and following the phytoplankton growth, the zooplankton starts to develop, namely herbivorous species reach the maximum of their abundance in May-June. Among those we mention phytoplankton filter feeders: Copepods, Appendicularia, Thaliacea, several small larvae.

Summer-Fall: After the break-down of the microalgae bloom in June-July, the abundance of zooplankton starts decreasing, to reach a minimum in August-September, coinciding with the stratification of water layers and the heavy thermocline which form a thermic barrier to the ascent of elements from the depth to the surface. Many groups became rare, whereas the meroplankton is enriched with various planktonic larvae of benthic organisms such as: Polychaeta Crustacea, Mollusca, Echinodermata and pelagic namely fish eggs and larvae, Crustacea and Cephalopoda larvae.

Exotic and Introduced species

36% of the dinoflagellates and 26% of diatom species present in the Levantine Basin including Lebanese sector, inhabit also the Red Sea; 45% of the planktonic fauna are common in the two marine Environments (Halim, 1969; Lakkis, 1980), most of them are considered as Lessepsian migrants (Table V). Migration process from the Red Sea into the Mediterranean, began after the opening of the Suez Canal in 1869. Previously to this period, few data on introduced species of fauna and flora of the area existed. More attention was given afterwards to the oceanography of the Eastern Mediterranean regarding the diversity and the impact of migration species (Por, 1978; Gurney, 1927).

Table V. Number of species in each planktonic group found in Lebanese waters. The number of exotic species, commonly found in the Red Sea and Levantine Basin.

Code	**Group**	**Nb.species in Lebanon**	**Nb.species in common with Red Sea**
DIA	Diatoms	160	40
DIN	Dinoflagellates	230	70
TIN	Tintinnids	141	40
FOR	Foraminifera	25	-
RAD	Radiolarian	25	?
HYD	Hydromedusae	74	11
SCY	Scyphomedusea	8	4
SIP	Siphonophores	28	18
PTE	Pteropoda	8	4
HET	Heteropoda	4	?
POL	Polychaetes & larvae	-	-
LCL	Cladocera	6	2
AMP	Amphipoda	25	7
COP	Copepoda	175	50
DEC	Decapod larvae	109	?
CHA	Chaetognatha	10	5
THA	Thaliacea	6	4
APP	Appendicularia	15	8
ICH	Ichtyoplankton	95	15

Inter-annual fluctuation

Variability between years of hydrology and plankton data are not very big; similar patterns in seasonal variations are observed from year to year depending on the parameters measured. The coefficient of variations for temperature between the years was 10% and for salinity 0.04% at the same stations. Inter-annual variability between the concentration of Chl.a was bigger than those for zooplankton biomass, respectively 25% and 15% (Fig. 9).

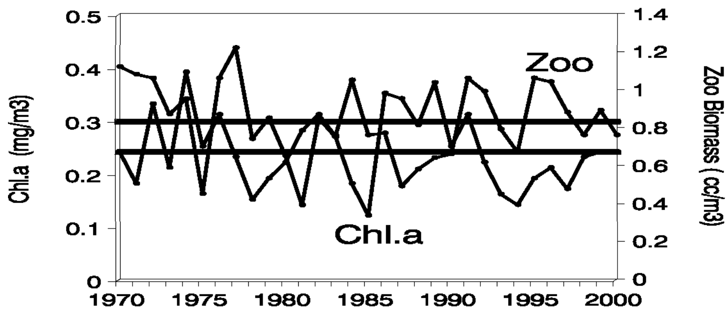


Fig. 9. Interannual fluctuations of Zooplankton biomass and Chl.a at surface seawater.

Discussion and conclusion

The eastern Mediterranean is a highly oligotrophic water body and a typical region of low productivity, due to limited nutrient supply in the euphotic zone (Ediger *et al.*, 1999). The hydrological and plankton time series data showed a clear regularity with minima and maxima, namely at surface layers. Seasonal variations of hydrological parameters and plankton in deep water are not as pronounced as at the surface. Inter-annual fluctuations are very similar year to year; on the other hand, they are more important between inshore and offshore waters. Vertical distributions of plankton data are more pronounced in the euphotic zone and neritic waters than in oceanic zone. These results are similar with those obtained in North East Mediterranean region, *e.g.* Iskenderoun Bay (Ylmaz and Tugrul, 1998). The extent of these variations is due mostly to hydro-climatic factors and nutrient cycles prevailing in the upper layers.

The water stratification and the heavy thermocline are more pronounced in the Levantine Basin than in other Mediterranean regions (Hecht, 1992). Vertical distribution of temperature is determined by the thermocline in summer and by isothermic conditions in winter; it has a great impact on vertical distribution of planktonic organisms. These spatial changes in plankton composition and abundance are significant between stations and between levels. Seasonal variation patterns of both phytoplankton and zooplankton are mostly similar between stations and between years, with a maximum in spring and a minimum late summer.

Over the 33 years of survey, we noticed certain changes in some hydrological parameters. During the last four decades we have observed 0.35‰, increasing salinity and 0.40°C temperature. This change is mainly due to the stop of the Nile flood after the building of the Aswan High Dam; the global warming might have some effect in this respect as well. During the last four decades, a certain "tropicalization" of the Levantine waters was observed (Lakkis *et al.*, 2003). This natural phenomenon is expressed by the increasing trend of the temperature and salinity that became close to those prevailing in the Red Sea. These hydrological changes have induced some ecological changes in the Levantine marine ecosystems, namely in the biodiversity following the Lessepsian migration of several marine species from the Red Sea to the Levantine basin.

Acknowledgements

This work was part of the Lebanese contribution to MEDAR/MEDATLAS II project, supported by European Union grant MAS3-CT980174. Thanks to IOC and OBI Conference Organizers for providing travel facility to attend the OBI conference in Hamburg.

References

Berman D.W, S.Z El-Sayed., C.C. Trees and Y. Azov. 1984. Optical transparency, chlorophyll and primary productivity in the eastern Mediterranean near the Israeli coast. Oceanologica Acta, 7(3):367-372.

Dugdale R.C. and F.P. Wilkerson. 1988. Nutrient sources and primary production in the Eastern Mediterranean. Oceanologica Acta, NS:179-184.

Ediger D., S. Tugrul, C.S. Polat, A. Yilmaz and I. Salihoglu. 1999. Abundance and elemental composition of particulate matter in the upper layer of northeastern Mediterranean.The Eastern Mediterranean as Laboratory Basin for the Assessment of Contrasting Ecosystems:241-266.

Goedicke T.R. 1972. Submarine Canyon on the central continental shelf of Lebanon. In: The Mediterranean Sea: Natural Sedimentation Laboratory D.J.Stanley Ed., 800p.

Gurney R. 1927. Cambridge Expedition to the Suez Canal 1924. Copepoda and Cladocera of the Plankton. Trans. Zool. Soc., London, 22:139-173.

Halim Y. 1969. Plankton of the Red Sea. Oceanogr. and Marine Biology. Ann.Rev., 7:231-275.

Hecht A. 1992. Abrupt changes in the characteristics of Atlantic and Levantine intermediate waters in the Southeastern Levantine basin 1992. Oceanologica Acta, 15(1):25-42.

Hecht A. and I. Gertman. 2001. Physical features of thye Eastern Mediterranean resulting from the integration of POEM data with Russian Mediterranean Cruises.Deep Sea Research, Part I, 48/8:1847-1876.

Krom M.D., S. Brennere, I. Israilov and B. Krumgalz. 1991. Dissolved nutrients, performed nutrients and calculatd elemental ratios in the south-East Medierranean Sea. Oceanologica Acta. 14(2):189-194.

Lacombe, H. and P. Tchernia. 1972. Caractères hydrologiques et circulation des eaux en Méditerranée, In: The Mediterranean Sea; a Natural Sedimentation Laboratory, edited by D.J. Stanley, Dowden, Hutchinson and Ross Stanley, 25-35.

Lakkis S. 1971. Contribution à l'étude du Zooplancton des eaux libanaises. Mar.Biol.11:138-148.

Lakkis S. 1980. A comparative study of the plankton in the Red Sea and Lebanese waters. Proc.Symp.Coast.Mar.Env. Red Sea,Gulf of Aden and Tropical W. Indian Ocean. (Khartoum)Vol.2:541-559. UNESCO-ALECSO.

Lakkis S. 1997. Long-time series of Hydrological and Plankton data from Lebanese waters (the eastern Mediterranean). NOAA Technical Report NESDIS 87:185-202.

Lakkis S. 2002. Archiving and Rescue of oceanographic data since 1965 in the Lebanese water (Eastern Mediterranean). In: Mediterranean and Black sea Database of Temperature and Salinity and Bio-chemical parameters Climatological Atlas. MEDAR/MEDATLAS II, 4CD, European commission (MAST) .

Lakkis S., A.E. Kideys, A.A. Shmeleva, E. Kovalev, E. Ünal and R. Zeidane. 2003. Comparison of Zooplankton biodiversity between Levantine basin and Black Sea, with reference to Alien species. Proceeding of the « Second International Conference on Oceanography of the Eastern Mediterranean and Black Sea: Similarities and Differences of two interconnected basins: 821-827.

Lakkis S. and V. Novel-Lakkis. 1981. Composition, annual cycle and species diversity of the Phytoplanckton in Lebanese coastal water. Journal of Plankton Research, 3(1):123-136.

Lakkis S., G. Bitar, V. Novel-Lakkis and R. Zeidane. 1996. Etude de la Diversité Biologique du Liban.Flore et Faune Marines. PNUE & Min. Agric. Beyrouth, Liban, Publ. No 6:123p.

Lakkis S., L. Jonsson, G. Zodiatis and D. Soloviev. 2003. Remote sensing data analysis in the Levantine Basin : SST and Chlorophyll-a distribution. Proceeding of the Second International Conference on Oceanography of the Eastern Mediterranean and Black Sea: Similarities and Differences of two interconnected Basins: 266-273.

Maillard C., M. Fichaut, H. Dooley and Medar/Medatlas Group. 2001. Medar-Medatlas Protocol, Part I: Exchange format and quality checks for observed profiles. R.INT.TMSI/IDM/SISMER/SISOO-084.

Malanotte-Rizzoli P., B. Manca, M.R. d'Alcala.and A. Theocaris. 1998. The Eastern Mediterranean in the 80's and in the 90's. The Big Transition Emerged from POEM-BC Observational Evidence. Proceeding of NATO Advanced Research Workshop on the eastern Medierranean as a Laboratory Basin for the Assessment of Contrasting Ecosystems. NATO Science Series; Kluwer Academic Publ.:1-7.

Parsons T.R. 1969. The determination of photosynthetic pigments in sea-water. A survey of methods. Determination of photosynthetic pigments in sea-water, UNESCO:19-37.

Por F.D. 1978. Lessepsian migration. The influx of the Red sea Biota into the Mediterranean by the way of the Suez canal. Ecological Studies. Spring Verlag, Berlin, N.York.

Souvermezoglou E. 1988. Comparaison de la distribution et du bilan d'échanges de sels nutritifs en Méditerranée et en mer Rouge. Oceanologica Acta, Océanographie pélagique mediterranéenne, ed. H.J.Minas et P. Nival, No SP:103-109.

Strickland J.D.H. and T.R. Parsons. 1972. A practical handbook of seawater analysis; 2nd ed. Bull. Fish. Res. Board Can., vol 167, 311p.

Townsend D.W., J. Christensen, T. Berman, P. Walline, A. Schneller and S. Yentsch. 1988. Near-bottom chlorophyll maxima in Southeastern Mediterranean shelf waters: upwelling and sediments as possible nutrient sources. Oceanologica Acta, No SP:235-244.

Yilmaz A. and S. Tugrul. 1998. The effect of cold and warm-core eddies on the distribution and stoichiometry of dissolved nutrients in the northern Mediterranean. Journal of Marine Systems:253-268.

TAXEX: TAXonomic EXpert system. History of development and technology of identification

Sergey Lelekov and Anton Lyakh

Dept. of Biophysical Ecology
Institute of Biology of the Southern Seas, Nakhimov av. 2, Sevastopol, 99011, Ukraine
E-mail: antonlyakh@gmail.com

Abstract

TAXEX is a series of taxonomic expert systems, which are developed to help scientists to professionally identify living organisms. They provide scientists with different taxonomic information, including taxon descriptions and diagnosis, geographic distributions, scientific nomenclature, identification keys and illustrations; it creates a tool for interactive identification of living organisms and trains new taxonomists. The main goal of TAXEX is to give public access to taxonomic and expert knowledge of the Black and Azov Sea biota. These systems can be used in interdisciplinary sciences like biological oceanography, biophysics, landscape ecology, bioecology, etc., in which specialists from different scientific fields are needed. Using taxonomic expert systems instead of high-paid taxonomists will reduce costs of scientific research and will allow many scientists without a specific biological education to work independently.

Keywords: TAXEX; Taxonomic expert system; Identification; Taxonomists training; Taxonomic knowledge base.

Introduction

The loss of biological diversity, our genetic heritage and the loss of habitats is accelerating in many parts of the world. That loss, exacerbated by our incomplete knowledge of the earth's biota, diminishes stewardship, restricts management, and imperils conservation of biological resources. Two components of this global problem are:

- loss in expertise necessary for the identification and inventory of biota
- poor state of knowledge of many aquatic and terrestrial organisms.

TAXEX (an abbreviation of TAXonomic EXpert system) is a series of taxonomic systems created to solve the problems mentioned above. They provide scientists with various taxonomic information, including taxon descriptions and diagnosis, geographic distribution, scientific nomenclature, identification keys, illustrations; it gives them a tool for interactive identification of living organism and allows training of new taxonomists. The main focus of the TAXEX Project is the Black Sea and Azov Sea region. The biodiversity of the Black and Azov Seas is well documented in local monographs and scientific papers, but there are no public Internet resources for those regions. In the past few years, interest in this region has increased, due to unique discoveries of species previously unknown to science, inhabiting the hydrogen-sulphide zone of the Black Sea (Sergeeva, 2003). Previous considerations stated that there was no life in that particular zone except for anaerobic bacteria. Besides traditional Black Sea species, some unique fish, which previously did not occur in the Black Sea, have now been observed, for example, *Sphyraena obtusata*, *Micromesistius poutassou*, and *Heniochus acuminatus*. New migrant species have been found and described as well. The creation of publicly accessible Internet resources will help enhance conservation of biodiversity of the Black Sea and Azov Sea region, it will enlarge our knowledge of life in the region and it will open a way to gather new information (Tokarev *et al.*, 2002).

Objectives

The objectives of the TAXEX Project are:

- to collect, describe and classify broad taxonomic resources of the Black and Azov Sea
- to maintain a vast knowledge of Black and Azov Sea biota in databases and knowledge bases and save it for future generations of scientists
- to store the knowledge of expert taxonomists in the right format within taxonomic expert systems
- to give public access on taxonomic and expert knowledge of the Black and Azov Sea biota to the scientific community and to fill the gap of non-accessible information about Black and Azov Sea flora and fauna
- to develop software for expert taxonomic identifications of the Black and Azov Sea biota, and for training new generations of taxonomists

History and enhancement of TAXEX

TAXEX is a taxonomic expert system – an interactive computer identifier of biological species – and a knowledge base including specific taxonomic information, a glossary of terms, references, etc. TAXEX has been developed since the end of 1980s at the Institute of Biology of the Southern Seas (Sevastopol, Ukraine). During the development of TAXEX, the algorithms to taxonomic identification and the interface to present this knowledge has changed since, but the traditional dichotomous taxonomic keys were never used. We tried to model the behavior of the expert, when he is identifying a taxon.

The first system's versions were working under MS DOS, but this had many limitations. The first version of the TAXEX Expert System was based on the conception of the frame – which is a computerized presentation of the expert's idea about the identified object. Defining the frame properties allowed to identifying taxa. These principles were the basis of a computer identifier of the Black Sea Isopoda (Lelekov *et al.*, 1996; Butakov *et al.*, 1997b). Drawbacks of this approach had become apparent under attempts to create new identifiers. The description of frames and the different rules on how to use them were so specific for every group of organism that the creation of a common method to forming frame descriptions became difficult. Hence, further developments were concentrated on the attempt to create a more universal model on the process of taxonomic identifications. Such a mechanism was developed (Lelekov, 1994) and used in computer identifiers for the Black Sea Fish Larvae, Black Seas Bivalvia, Black Sea Gastropoda, and Fishes of the Black and Mediterranean Seas (Butakov *et al.*, 1996; Butakov *et al.*, 1997a; Butakov *et al.*, 1998). These expert systems work under MS Windows and manage taxonomic knowledge that is stored in a database.

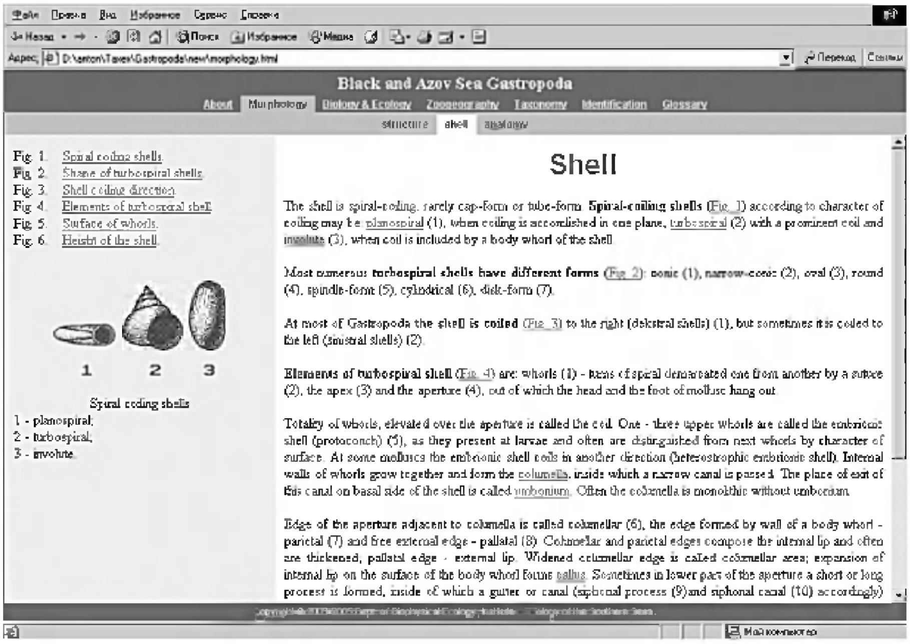


Fig. 1. TAXEX System of the Black and Azov Sea Gastropoda.

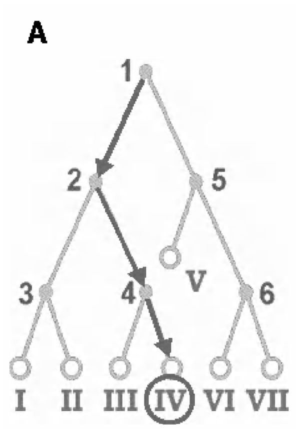
Providing public access to taxonomic and expert information needed the creation of a new generation of online software tools, which can work on both the Intranet and Internet. The Java version of our TAXEX Expert System was developed at the end of the 1990s, and now is a Java applet that uses information stored in identification tables, and the taxonomic knowledge base consisted of a set of linked HTML pages (Fig. 1). TAXEX can now be easily distributed and it is accessible for everyone.

Identification algorithms

To describe the identification algorithms used in TAXEX we first considered a classical taxonomic identification scheme based on a dichotomous key. It can be presented as a binary tree, where the nodes contain the descriptions of the taxonomic characters, and the leaves contain the description of taxa. To identify an organism it is necessary to consecutively traverse the identification tree from its root till one of the leaves (Fig. 2, A). For every step you need to choose the state of a character appropriate to your taxon and select the next direction of movement.

This dichotomous identification process has the following disadvantages:

- at every step of the identification you can only choose between two variants that increases the number of necessary steps
- the path is fixed: you move from tree up till a tree leaf and no character can be omitted
- the states of some characters are undefined for some taxa, if the characters and their taxa lie in different branches
- if you cannot find the state of a character, when an organism for instance is damaged, further diagnosis becomes impossible.



B

	1	2	3	4	5	6
I	+	+	+			
II	+	+	–			
III	+	–		+		
IV	+	–		–		
V	–				+	
VI	–				–	+
VII	–				–	–

Fig. 2. (A) Representation of a classical identification key in the form of a binary tree. Arabic numerals are identification characters. Roman numerals are identified taxa. Arrows show one of the possible ways of identification. (B) Identification table, which corresponds to identification tree. The arrow above of the table shows the path of diagnosis.

Identification trees can be converted into an identification table – a matrix, where the rows contain the states of all the identification characters for a given taxon, and the columns contain the state of a given character for every taxon. Using an identification table, the organism's identification procedure can be presented as the consecutive division of the group of taxa into subgroups while a minimal element – a taxon – would be reached. The path of diagnosis is determined by the order of columns (Fig. 2, B).

The identification table for a binary taxonomical key has many empty cells, and the filled cells can only contain two different values, + or – (yes or no). To improve this situation we propose to use a new identification table, called improved identification table in which:

- there are no empty cells, that is: no undefined characters
- characters can have more than two states
- cells can keep more than one character state.

We have built our identification process on the improved identification table that allows any user of the TAXEX System to fill in the cells more freely, when he/she is diagnosing an organism. At every step of the identification, the user has to answer the questions about some taxon characters, and he can:

- omit the character, if he could not determine its state, for example when the organism is partially damaged
- choose more than one answer, when he/she notices that the taxon in question has multiple character states, or he/she is not sure of the accuracy of the character determination.

To identify a taxon in such non-rigorous conditions, the TAXEX System operates by a hypothesis concerning the taxonomical position of the organism. In the beginning all taxa have maximal probabilities, and the system supposes that the diagnosed taxon belongs to the higher taxonomic rank kept in the system knowledge base. Through the selection of states, the probability of taxa that do not have selected states decreases. (Fig. 3) Note, that probabilities of the hypothesis can be decreased only.

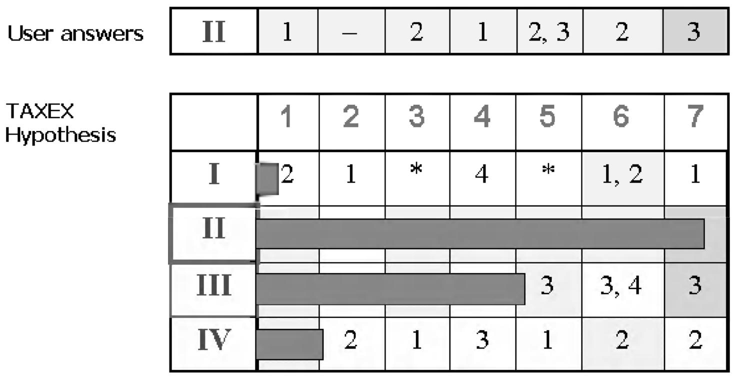


Fig. 3. Hypothesis of TAXEX after a user has answered the system's questions. Above are the states of the characters of the identified object. Below is the advanced identification table with horizontal bars, which shows the TAXEX hypothesis. In this case, the most probable hypothesis is that the identified object is taxon II.

When the identification is finished, the user obtains the taxon with the maximal probability as the result of the diagnosis. Certainly, he can view all other system versions with lower probabilities. If the user knows the taxonomic rank of the organism, he can alter the current system hypothesis, selecting the name of order, family, genus, etc., this allows omitting questions concerning other taxonomic groups and reducing the number of identification steps.

To reduce the number of necessary steps TAXEX tries to choose that character which could confirm the most probable hypothesis. This character has to divide the group of the most probable taxa into subgroups. At every step TAXEX tries to choose the best dividing character according to the current hypothesis and remaining not identified characters. Moreover the best divider has to satisfy on two criteria:

- the first criterion takes the probability of the occurrence of a taxon into account. Some species are common species. You observe them very often, and sometimes are present in every sample you study. Other species are rare, you only encounter them in few samples from limited regions, or you almost never find them. Therefore, in the beginning, it is useless to ask the user about the characters of rare species. Firstly, the system supposes that it is a common taxon and tries to identify it. When the diagnosis does not give any appropriate results, the system will ask you about the rare species characters
- the second criterion takes into account the cost of taxon characters determination. Usually, characters with small determination cost do not need special instruments; they are external and can be easily distinguished. Characters with large determination costs usually are internal; you need special instruments, such as a microscope or scalpel to examine them. Therefore, TAXEX first asks questions with small determination costs and gives you the possibility to identify a taxon without additional operations, like microscopic observations or dissections.

Costs of character determinations and probabilities of sampling a species in nature are set up by taxonomists, who developed the identifier. The best divider is chosen by the system at each step of identification.

As a result a new version of TAXEX identifiers has been developed. Its functioning is based on the following information:

- identification table, which set the correspondence between taxa and their characters
- costs of character determination
- probability of sampling of a certain taxon in nature.

The identification information is stored in a knowledge base, which also keeps data on species and taxonomic group descriptions, biology, ecology, biogeography, drawings and photographs, glossary of terms, bibliography, etc. The knowledge base and computer identifiers together constitute the Taxonomic Expert System.

Training

Training of object identification consists of the following elements:

- studying the relation of specimens to all object classes
- the study of distinctive characters of specimens from different classes
- becoming familiar with identification procedures.

Instructors with excellent expertise and good quality training capabilities, including access to biological collections, atlases of animals and plants, identification keys, are necessary. For young universities, where scientific schools are just about getting started, the availability of both the instructors and the training equipment is often a problem. In these cases, TAXEX can help to solve this lack of resources. TAXEX Expert Systems allow accomplishing professional object identification, and they have special tests for the teaching and training of users. Any expert system includes four tests:

- How good do you know a specimen? The test is used to develop a pupil's ability to identify a specimen by using specific characters. During the test the pupil has to select the correct states of these characters.
- How good do you know taxon characters? This test is used for training the pupil to distinguish taxa within a taxonomic rank. The system shows characters, points out their states and asks the pupil to choose the taxa to which these characters are related.
- Determine a taxon by its characters. This test is used to gain knowledge about an identified taxon. During the test the TAXEX system enumerates states of the taxon characters and at the end asks the pupil to indicate the taxon, to which these characters are related.
- Determine a taxon by images. The test is used to train the pupil's ability to recognize a taxon by the use of images. The system shows one or more images of possible taxa and the pupil has to indicate its taxonomic name.
-

These four tests can serve as a good methodical basis for preparing new specialists. Moreover, interactive training tools stimulate the pupil to actively learn how to solve questions, hereby making use of the TAXEX knowledge base, in addition to other sources of information and expert knowledge. One of the important advantages of the system is the possibility to put it on the Internet. This allows organizing of distance learning.

Conclusion

By using TAXEX you will get access to expert knowledge and will be able to identify taxa like an expert. These systems can be used in interdisciplinary sciences like biological oceanography, biophysics, landscape ecology, bioecology, etc., in which specialists from different scientific fields are needed. Using taxonomic expert systems instead of high-paid taxonomists will reduce the costs of scientific research and will allow many scientists without a specific biological education to conduct their research more independently.

Identification tools are also useful for young taxonomists, who just begin to learn to identify species. Four tests, included in the expert system, can serve as good methodical basis for preparing new specialists. Interactive training tools stimulate the pupil to actively learn how to solve questions, hereby making use of the TAXEX knowledge base, other sources of information and expert knowledge. One of the important advantages of the system is the possibility to put it on the Internet, which allows organizing of distance learning.

References

Butakov E.A., S.G. Lelekov and V.D. Chuhchin. 1996. Kratkiy opredelitel dvustvorchatyh molluskov Chernogo moray treh nadsemeystv *Cardioidea*, *Veneriidea*, *Tellinoidea*. [Brief identifier of Black Sea molluscs of superfamilies *Cardioidea*, *Veneriidea*, *Tellinoidea*]. Institute of Biology of the Southern Seas. Sevastopol. 46p.

Butakov E.A., S.G. Lelekov and V.D. Chuhchin. 1997a. Opredelitel bruhonogih molluscov *Gastropoda* Azovo-Chernomorskogo basseyna. [Identifier of Black and Azov Sea *Gastropoda*]. Institute of Biology of the Southern Seas. Sevastopol. 127p.

Butakov E.A., S.G. Lelekov, E.B. Makkaveeva and V.F. Zhuk. 1997b. Opredelitel rakoobraznih otryada *Isopoda* Chernogo morya. [Identifier of Black Sea *Isopoda*]. Institute of Biology of the Southern Seas. Sevastopol. 79p.

Butakov E.A., E.Yu. Georgieva, S.G. Lelekov, N.P. Pakhorukov and L.P. Salekhova. 1998. Opredelitel semeystv ryb Sredizemnogo morya. [Identifier of Black Sea fish families]. Institute of Biology of the Southern Seas. Sevastopol. 483p.

Lelekov S.G. 1994. K voprosu vybora posledovatelnosty priznakov v diagnosticheskih expertnih systemah. [To the question of choosing characters in diagnostic expert systems]. Kibernetica 4:167-173.

Lelekov S.G. 1995. PC-based identification of species for ecological monitoring. p.179-183. In: ECOSET'95. International conference on ecological systems enhancement technology for aquatic environment. The Sixth International Conference on aquatic habitat enhancement. Tokyo.

Lelekov S.G. and E.B. Makkaveeva. 1996. Komputernoye opredeleniye rakoobraznih otryada *Isopoda* Chernogo morya. [Computer identification of Black Sea order *Isopoda*]. Zoological Journal 75:35-44.

Sergeeva N.G. 2003. Meiobenthos of deep-water anoxic hydrogen sulphide zone of the Black Sea. p.880-887. In: Oceanography of the Eastern Mediterranean and Black Sea. Similarities and differences of two interconnected basins. Yilmas A. (Ed.). Tubitak Pube.

Tokarev Yu.N., S.G. Lelekov, V.V. Melnikov and A.M. Lyakh. 2002. Perspectivy ispolzovaniya computernih opredeliteley v oblasty taxonomii. [Perspectives of using computer identifiers in taxonomy]. Ecologiya moray 61:95-98.

A Semantic Modelling Approach to Biological Parameter Interoperability

Roy Lowry[1], Laura Bird[1] and Pieter Haaring[2]

[1]British Oceanographic Data Centre Joseph Proudman Building, 6 Brownlow Street, Liverpool L3 5DA, UK
E-mail Roy Lowry: rkl@bodc.ac.uk, E-mail Laura Bird: labi@bodc.ac.uk

[2]Ministry of Transport, Public Works and Water Management Directorate-General of Public Works and Water Management (Rijkswaterstaat). National Institute for Coastal and Marine Management (RIKZ) P.O. Box 20907, 2500 EX, The Hague, The Netherlands
E-mail: p.a.haaring@rikz.rws.minvenw.nl

Abstract

The BODC Parameter Dictionary currently contains over 16,500 terms of which nearly 11,000 pertain to biological parameters. The Rijkswaterstaat database in the Netherlands covers over 10,000 types of measurement, most of which are either chemical or biological. A requirement to populate a metadatabase described in terms of the BODC dictionary from the Rijkswaterstaat database meant that parameter interoperability between these information sources needed to be addressed. One technique for approaching this is manual mapping, working term by term through one of the information sources then searching for matching terms in the other. However, whilst this may be feasible for dictionaries containing tens of terms, it is totally unrealistic when the counts run into the thousands and so an alternative, automated approach was required.

Automation was initially attempted using a semantic matching tool developed at Rijkswaterstaat to offer a restricted list of BODC terms (preferably a single term) as the possible matches for each measurement. However, this met with limited success because the BODC dictionary consisted of plain language terms that not been written with machine processing in mind and had no constraints on either syntax or vocabulary. To appreciate the problem consider the programming required to recognise that '*Calanus* abundance', 'Number of *Calanus*', '*Calanus* count' and 'Abundance of *Calanus*' essentially mean the same thing. Further, no dictionary, especially a dictionary without vocabulary constraints, is perfect and there is a high risk that matches will be missed due to basic errors such as spelling mistakes.

The Rijkswaterstaat database is described in terms of a data model that qualifies measurements through associated attributes describing, amongst other things, what was measured and how it was measured. This is an example of a semantic model in which an entity is described in terms of discrete items of information, called semantic elements. Ideally, these elements are atomic, unambiguous and therefore ideally suited to machine interpretation. It was concluded that the only way a mapping could be achieved would be to develop a model along similar lines to describe the BODC dictionary and then map the two models.

A prototype semantic model based on three sub-models, each containing between 10 and 12 semantic elements, was developed to describe the biota, biota composition and chemical terms in the BODC dictionary and populated with approximately 13,000 terms. This was used as a basis for a two-stage mapping to the Rijkswaterstaat data model. The first stage was to set up a mapping

between the semantic elements in the two models. For example, it was established that the 'Parameter' element in the Rijkswaterstaat model was equivalent to the concatenation of the 'Param' and 'Param_Comp' elements in the BODC semantic model. The second stage was to produce a mapping between the vocabularies used in each set of matched semantic elements. For example, the Rijkswaterstaat compartment term 'Surface water' mapped to the BODC compartment term 'water column'. Once these mappings had been established an automated term generation procedure was used to translate sets of Rijkswaterstaat semantic elements into BODC terms and identify matches.

The result was an automated mapping for approximately 90% of the Rijkswaterstaat measurement description terms. Of the remainder, most were matched by straightforward extensions to the vocabulary mapping. However, a small number of problems remained that could only resolved by querying Rijkswaterstaat, including ambiguity caused by homonyms that only came to light through standardisation of the BODC model to the ITIS taxonomic database.

This exercise has shown that semantic modelling is a very promising technique for automating parameter interoperability between biological databases. However, without standardisation, particularly in the description of taxonomic entities, matches will be missed and there is a small but significant risk of false matches between parameters that are totally different.

Keywords: Parameters; semantic modelling.

Introduction

This paper documents the work done to develop parameter interoperability between the biological and chemical data holdings of the British Oceanographic Data Centre (BODC)[1] and the Dutch Rijkswaterstaat[2].

Description of the Problem

BODC and Rijkswaterstaat have large marine databases holding a wide range of physical, chemical and biological measurands (the Open GIS Consortium (OGC) term for something that has been measured). Both organisations participated in two EU projects, EDIOS[3] and SEA-SEARCH[4] that developed pan-European metadatabases, which use a measurand discovery vocabulary[5] developed by BODC. As part of the vocabulary development, a mapping was built to the BODC measurand mark up (the BODC Parameter Usage Vocabulary[6]) but no such mapping existed for the Rijkswaterstaat measurand mark up. There were two possible solutions to this problem:

- Mapping the Rijkwaterstaat measurand mark up and the BODC discovery vocabulary
- Mapping between the measurand mark ups of the two organisations, allowing the BODC discovery vocabulary mapping to be used.

It was realised that whilst the second approach was more difficult and would involve more work, especially enhancement of the BODC Parameter Dictionary, the resulting parameter interoperability offered significantly greater reward. Resources for BODC dictionary development were available through the NERC EnParDis[7] (Enabling Parameter Discovery) project. Consequently, the full mapping approach was taken.

Measurand Mark up in BODC and Rijkswaterstaat

There are significant differences between the measurand mark up strategies used by BODC and Rijkswaterstaat. The BODC system has its roots in the GF3 data model[8] in which measurand instances are linked to a key (termed the parameter code) defined by an entry in a parameter dictionary. This specifies one or more items of information about what the measurand is and how it was obtained. As the BODC database expanded from physical into biological and chemical data, limitations of legacy data formats resulted in a significant increase in the parameter code information load.

The Rijkswaterstaat database was designed around the DONAR[9] data model. This has the measurement as the primary entity, which is linked to a set of attributes containing specific atomic items of metadata describing the measurand and where, when and how the measurement instance was made. Each item of metadata is populated from a controlled vocabulary. The DONAR Parameter Dictionary is therefore simply a catalogue of valid combinations of metadata information items pertaining to the identity of the measurand and how it was made linked to a key.

The Starting Position

At the start of the mapping exercise in October 2003 the BODC dictionary described the mark up code through two plain text fields containing up to 200 bytes each. These had been populated over the 25 years of the dictionary's development in a less than consistent manner. Certain information categories were sometimes in one field and sometimes in the other. The grammatical structures were inconsistent and consequently the fields could not be concatenated sensibly. Whilst this situation was acceptable for interpretation by a human, it was totally inadequate for use by software agents.

In contrast, the DONAR presented Rijkswaterstaat with a particular type of information in a consistent and readily identified field within the data model. Furthermore these items of information, which we will term semantic elements, could be concatenated to provide comprehensible measurement descriptions in both Dutch and English. The system could therefore be used both by software agents and for presentation through a user interface.

Dictionary Mapping

At its most basic level, the mapping between the DONAR catalogue and the BODC Parameter Dictionary involves the following steps:

- For each entry in the Rijkswaterstaat catalogue (delivered as a spreadsheet with one column per semantic element):

- Use BODC dictionary search tools (Microsoft Access Filter by Form) to locate the entry having the same meaning as the combination of semantic elements
- If found: Copy code from Access form and paste into the DONAR spreadsheet
- Else: Manually prepare a dictionary entry record and submit to for quality assurance and loading.

This process is tedious, error prone and pushes the limits of human endurance for dictionaries with more than a couple of hundred entries. This mapping exercise involved thousands and it became obvious that a mechanism for automating the procedure was required.

The first attempt at automated mapping was developed by Pieter Dekker (from Xi advise bv) and used semantic analysis on the two BODC dictionary plain text fields to identify the DONAR element combinations that were a close match. The user then selected which one to use. This failed to work in practice for two reasons. First, the system had no mechanism to expand the population of the BODC dictionary. Consequently, if the Rijkswaterstaat element combination wasn't covered, there was no way in which it could be mapped. Secondly, the vocabulary and syntactic structure of the BODC dictionary plain text fields were not standardised. Human intelligence can recognise that '*Calanus* abundance', 'abundance of *Calanus*', 'number of *Calanus* per unit volume' have the same meaning, but artificial intelligence cannot without an extensive domain thesaurus or ontology.

BODC Dictionary Development

Following a presentation of the DONAR model and the semantic mapping tool at Rijkswaterstaat in December 2003, it became apparent that the BODC dictionary required drastic improvement if the mapping was to succeed. The approach taken was based on the extension of the DONAR design principles to the scope covered by the BODC Parameter Dictionary. In particular, the ability to combine semantic elements into meaningful text descriptions was enhanced.

The DONAR model per se was not adopted because there was already evidence in its usage at Rijkswaterstaat of 'shoehorning' where multiple items of information were forced into a single semantic element because they were needed and there was nowhere else for them to go. The scope of the BODC dictionary would make the problem much worse. For example, some BODC zooplankton data includes development stage information that would have had to be included in the same element as the taxon name.

The model developed for biological dictionary entries currently contains the following semantic elements:

- Parameter (Abundance, Biomass)
- Taxon_code (Integrated Taxonomic Information System (ITIS[10]) code)

- Taxon_name
- Taxon_subgroup (gender, size, stage)
- Parameter_compartment_relationship (per unit volume of the, per unit area of the)
- Compartment (water column, bed, sediment)
- Sample_preparation
- Analysis
- Data_processing.

Element content is governed by a controlled vocabulary, with any elements that are not relevant to a particular dictionary entry coded as 'not specified'.

The elements may be combined into text descriptions like:

'Carbon biomass of Urotricha (ITIS 46243) <20um per unit volume of the water column by optical microscopy and abundance to carbon conversion using the equation of Putt & Stoeker (1989)'

This is one of three sub-models currently being developed to cover the scope of the BODC dictionary, the other sub-models being for contaminant in biota data and a 'chemical' sub-model that seems to cover everything except biology. Once the model population has been completed these sub-models will be combined into a single element superset. Further atomisation of the model will also be undertaken at this stage where 'shoehorning' has been observed, such as division of the biological sub-model taxon_subgroup element into 'gender', 'size', 'development stage' and 'taxon_subgroup'.

Semantic Model Mapping

Mapping between two semantic models is a two-stage process. The first stage is to produce a mapping between the semantic elements in the two models. For example, the DONAR 'parameter' semantic element contains entries such as 'biomass per surface area unit' and 'number per volume unit', which are concatenations of the BODC model elements 'parameter' and 'parameter_compartment_relationship'. Note that it is by no means certain that this mapping will be a simple one-to-one relationship, particularly if shoehorning has occurred during population of the model instances.

The second stage is to produce a mapping between the vocabularies for the mapped elements. For example, the DONAR 'compartment' element maps to the BODC element of the same name. A subset of the vocabulary map is as follows:

Rijkswaterstaat	**BODC**
soil/sediment	bed
suspended solids	suspended particulate material
surface water	water column
porous water	sediment pore water

This two-stage process normalises the mapping procedure, cutting down the number of comparisons required by at least an order of magnitude. Furthermore, as the semantics of the element vocabularies are simple, mapping automation becomes an achievable goal. In practice, once BODC dictionary population issues had been addressed, over 90% of the final map was achieved by running a single SQL statement. Manual expansion of the vocabulary maps to deal with instances of different names meaning the same thing brought the level of completion to over 99%. Develop of thesaurus servers would allow automation of this part of the process as well.

There were a small number of DONAR combinations that required significant manual effort to achieve a mapping due to unclear or ambiguous semantics in the element vocabularies. For example, after an e-mail exchange to clarify semantics the term 'residual beta' was mapped to 'beta emitters other than 3H and 40K'. It would therefore seem that fully automated mapping is currently not an achievable goal.

An obvious, but important, point is that if a complete map is to be produced then one of the models must be the superset of the other. Until the Nirvana of an all-encompassing model is achieved, this will inevitably mean that the population of one of the models will need to be expanded as part of the mapping exercise. Adding records as part of a manual mapping exercise is a long and tedious process. However, the dictionary expansion requirement from semantic model mapping involves either adding new combinations of existing vocabulary members or a vocabulary extension. This was achieved quickly and relatively easily for the mapping exercise documented here in a semi-automated manner using a basic general-purpose tool (Oracle's SQL*PLUS). Bespoke tools are currently under development that will make the job easier still.

Checking the map produced revealed some errors due to homonyms such as Branchiura. The taxon identifier fields used contained names with no qualifying information such as a reference or a taxonomic database key. The exercise emphasised that this standard of labelling is insufficient to support totally reliable automated interoperability between biological databases. The BODC semantic model now includes an ITIS key element as a result of the lessons learned.

Semantic Model versus Parameter Dictionary

It is clear from this exercise that the semantic model is a much more powerful interoperability tool than the parameter dictionary (a vocabulary describing measurands through a plain text description). Furthermore, mapping measurands across databases uses only part of their potential. The map generated in this exercise can be used to determine that a given measurement in the Rijkswaterstaat database is exactly the same thing as a measurement in the BODC database. Such measurements may obviously be safely combined into a composite data set. However, what about cases where measurement descriptions are nearly the same, or where 'fit for purpose' criteria determine the measurements that may be safely combined?

Consider the following example for chlorophyll. The following are two entries from the BODC dictionary generated by concatenation of elements from the chemical semantic sub-model:

Concentration of chlorophyll-a {chl-a} per unit volume of the water column [particulate >30um phase] by filtration, acetone extraction and fluorometry

Concentration of chlorophyll-a {chl-a} per unit volume of the water column [particulate 0.6-5um phase] by filtration, acetone extraction and fluorometry

A user wishing to determine the timing of the spring bloom would happily merge data corresponding to both of these descriptions. A user with a compartmentalised biogeochemical model would not as the two measurements represent two completely different phytoplankton communities. If our first user was given control over which semantic model elements were used to build the description both the above could be reduced to:

Concentration of chlorophyll-a {chl-a} per unit volume of the water column

In this way we have provided a simple mechanism for user control over the scope of data interoperability. This same mechanism could also be used as the basis for a data set discovery interface. By turning elements on or off the user is able to control the level of information detail used in the subsequent search. Note that in this scenario it is not always necessary to specify precise semantic elements values for a search. In many cases it will be sufficient to simply say that the element should be other than 'not specified'.

It can therefore be seen that semantic models provide far more than just a tool for automated parameter mapping.

Conclusions

The following conclusions were drawn from this work and other associated activities in the EnParDis project:

- Manual mapping of measurand metadata is only feasible on the smallest of scales and that automation is not possible if the metadata is encoded as unstructured plain text
- Automated mapping becomes feasible if the description is encoded as atomised semantic elements, but it could be further improved by the availability of domain-specific thesauri and ontologies
- Standardisation of vocabularies, especially the utilisation of keys from published reference sources, renders automated mapping both easier and more reliable
- 99% of a map is completed in 10% of the time required to produce a complete map
- Semantic models may be used for purposes other than parameter mapping

- The conversion of the 16,500 term BODC parameter dictionary from plain text descriptions to a semantic model has significantly enhanced its value as a tool for database federation and data interoperability.

Acknowledgements

The authors are indebted to many colleagues in their parent organisations, the international marine XML community (the ICES/IOC SGXML group and the EU MarineXML project) and the MMI project for stimulating discussions that have helped shape this work and for the encouragement that has helped it progress. The work at BODC was supported financially by the UK Natural Environment Research Council through the EnParDis project.

Notes

[1]British Oceanographic Data Centre: http://www.bodc.ac.uk
[2]Rijkswaterstaat: http://www.verkeerenwaterstaat.nl/?lc=uk or
http://www.rijkswaterstaat.nl (Dutch language)
[3]European Directory of the Ocean-observing System: http://www.edios.org/
[4]SEA-SEARCH: http:// www.sea-search.net
[5]The BODC Parameter Discovery Vocabulary:
ftp.pol.ac.uk/pub/bodc/jgofs/datadict/new/parameter_group.csv
http://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/
[6]TheBODC Parameter Usage Vocabulary:
ftp.pol.ac.uk/pub/bodc/jgofs/datadict/new/parameter.csv
http://www.bodc.ac.uk/data/codes_and_formats/parameter_codes/
[7]Enabling Parameter Discovery (EnParDis): http://www.bodc.ac.uk/projects/uk/enpardis/
[8]GF3, A General Formatting System for Geo-Referenced Data, IOC Manuals and Guides 17, UNESCO 1987
[9]DONAR: http://www.waddensea-secretariat.org/news/symposia/Demowad/RIKZ.html
[10]Integrated Taxonomic Information System: http:// www.itis.usda.gov

Databases as a tool for studying the dynamics of macro- and meiobenthos on algal communities in the Black Sea

Mazlumyan S.A. and E.A. Kolesnikova

Institute of Biology of the Southern Seas, NASU, Sebastopol, Ukraine
E-mail: mazl@sevpochta.com.ua

Abstract

As an ecosystem component, meiobenthos contributes to the condition and sustainability of the ecosystem. Meiobentic species are characterised by having a fast response and a short life cycle, they have high energy flow rates and are involved in biochemical processes. Meiobenthic response varies in a wide range, but on the other hand, it is rather predictable, and that makes meiobenthos a useful tool in monitoring. The interannual observations revealed that benthos developed fast responses to unfavourable conditions in 1990 and this makes meio- and macrobenthos a dynamic but vulnerable ecosystem component. The database was created using a seasonal index, determined for the main taxonomic groups of benthos that are associated to algae; comparative analysis was carried out to understand how the index varied during 23 months for identical taxa occurring in three different localities.

Keywords: Black Sea; macro- and meiobenthos; macrophytes; short- and long-term variations.

Introduction

There are huge stocks of data available on coastal algal communities and their associated species, inhabiting the contouring biotopes of our world oceans. In a favorable environment, the fauna occurring on macrophytes is diverse and the density of macro- and meiobenthic species is high.

First qualitative studies of Black Sea meiobenthos go back to the 19th century, quantitative studies to 1955. The creation of a database on taxonomy and the distribution of benthic organisms is essential because the dimension of relevant faunistic information is large. A benthic sample may contain more than ten different taxa and each taxonomic group may count for about 1,000 individuals. According to different methods, samples can be collected in several replications (up to 5 samples). This also adds to the dimension of data. In addition, increasing the geographical and temporal scale of investigation leads to an increasing dimension of the data matrix.

The database has been created using a taxonomic time series (35 years) from a geographically restricted area, more specifically a bay in south-western Crimea. The study focused on the communities found on the algae *Cystoseira crinita* in Kruglaya bay.

Materials and Methods

The three sampling sites with different environmental conditions are located in: the open coastal water (I), at the mouth of the bay (II) and amidst the bay (III). At each sampling site, the samples were collected at 0.7m depth during each month from April 1990 to March 1992. In each sample, the fouling organisms were counted, identified and their numbers determined per 1 kg algal wet weight. Altogether, 190 samples were collected and handled, and ten taxa: Turbellaria, Nematoda, Polychaeta, Acarina, Harpacticoida, Ostracoda, Amphipoda, Bivalvia and Gastropoda were examined.

The database on the taxonomy and distribution of meiobenthos proved to be an efficient tool in addressing some aspects of biodiversity, such as interannual, seasonal and long-term variations. K-dominance analysis was carried out. All the variations were studied in relation to each site.

Results

Seasonal dynamics

Using the database on taxonomy and distribution of meiobenthos in the samples, the trends of seasonal dynamics were obtained for macro- and meiobenthic organisms associated to algae (Fig. 1). Examination of the variations shown in the main taxonomic groups pointed out that for Acarina at locality I and II, the peaks developed concurrently in summer and in winter. In locality III the peaks developed in autumn and in spring; the only peak found in winter concurred with the peak in the 1st locality (Fig. 2). The diversity of peaks for Acarina is due to interspecific variations in reproduction time. Moreover, some species live in soft bottom sediment during almost all year round, while there is much more variation for species depending on short-term algal growths.

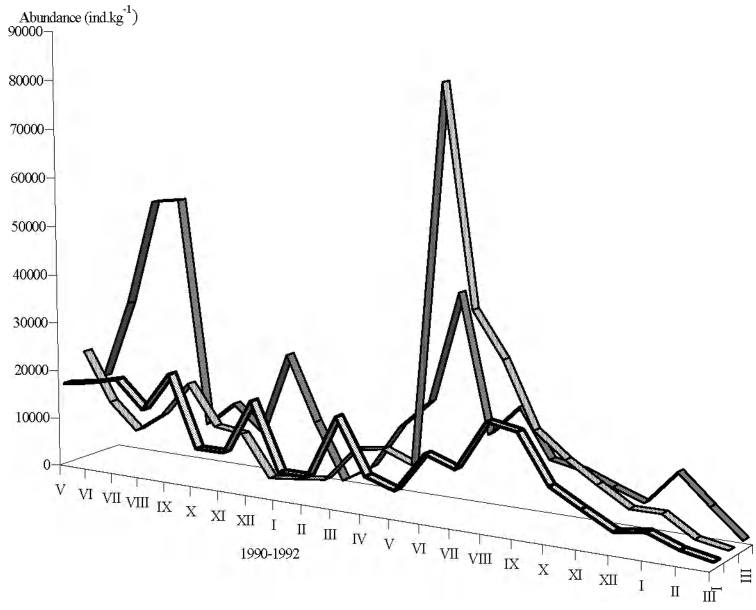
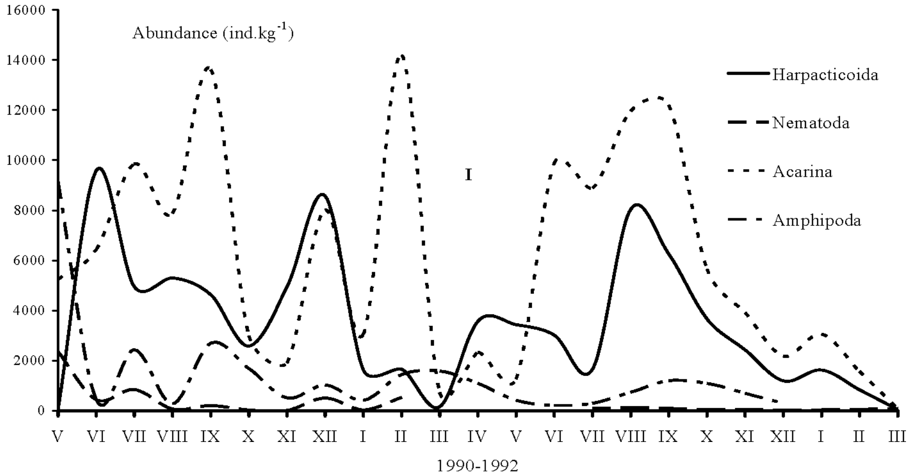
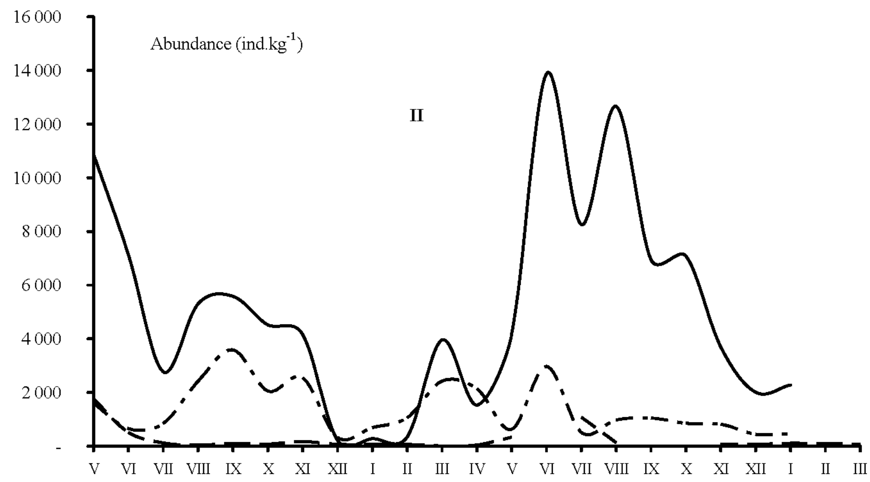


Fig. 1. Seasonal dynamics of the macro- and meiobenthic associated to Cystoseira crinita.





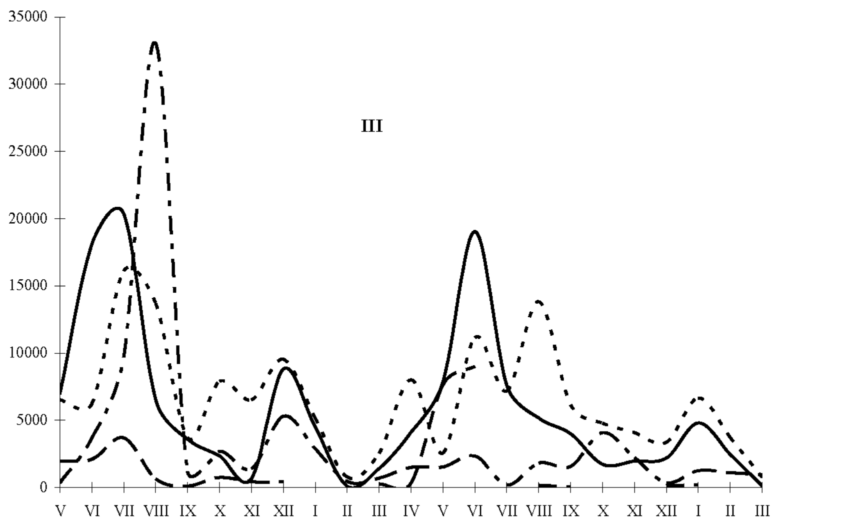


Fig. 2. Seasonal dynamics of meiobenthos taxa in the Kruglaya Bay at points I, II, III.

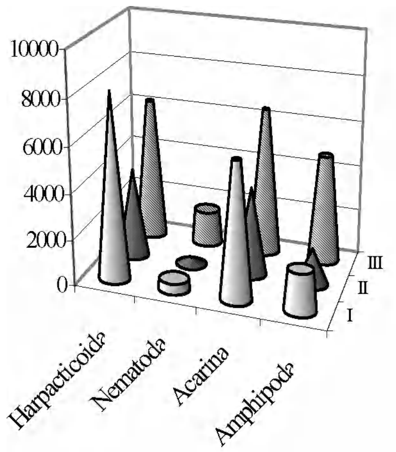
For Harpacticoida, winter peaks developed concurrently in all the three localities in 1990 and 1992, and summer peaks were present in 1990 and 1991 (Fig. 2). A winter peak was not found in 1991. Seasonal dynamics in abundance of Harpacticoida is related to periodic reproduction. Phytophilous harpacticoids have two reproductive strategies: perennial and seasonal. Species with an all-year-round reproduction also have seasonal peaks of reproduction during different seasons.

For Amphipoda, summer peaks were simultaneously found in all three localities in 1990; in 1991 autumn peaks concurred for the 1st and the 2nd localities (Fig. 2). During the full period of investigation the 3d peak was registered in the 1st, 2nd and 3rd localities in winter, summer and autumn. In the biocenosis of *Cystoseira crinita*, the biomass and population density of amphipods have two peaks, one in spring (April) and the other in summer (July-August). These seasonal peaks are directly depending on the biology of amphipods: their biomass reaches a maximum in spring, when many post-winter survivors reproduce. In summer, the peak in population density is related to the reproduction of juvenile individuals of the first generation, while the biomass is considerably lesser than it was in spring.

Interannual variations

Using the taxonomic database, statistical characteristics were calculated for all meiobenthic taxa, associated to algae for each of the three studied localities during 12 months arbitrarily selected between1990 and1992. This approach has clarified interannual changes that were typical for the main taxa.

During our study, the marine organisms were unevenly distributed, both quantitatively and qualitatively (Fig. 3, 4).



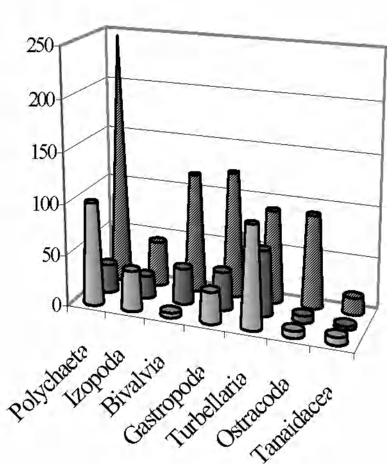


Fig. 3. Average abundance (ind.kg^{-1} algal wet weight) for the major meiobenthos taxa in the Kruglaya Bay atlocality I, II, III (May 1990 - May 1991).

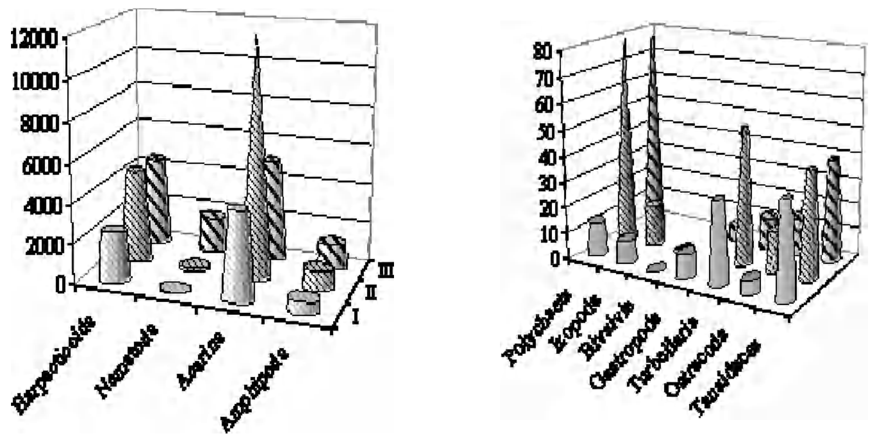


Fig. 4. Average abundance (ind.kg^{-1} algal wet weight) for the major meiobenthos taxa in the Kruglaya Bay at locality I, II, III (March 1990 - March 1991).

In places where the algae occur and where the activity of waves is relatively low, the population density was higher than in the open sea. Taxa are also distributed unevenly. Nematodes were predominating in the protected area of the bay, the Acarina in the mouth of the bay and harpacticoids were abundant at all localities. Sometimes Amphipods were plentiful in some samples and absent in others, although collected at the same site. Amphipods were numerous in the plankton above the algal growth. Isopods, turbellarians and polychaetes were found all year round but in small numbers, and gastropods and juvenile bivalve molluscs only during spring and summer.

Analysis of the structure of the investigated community pointed out that Acarina and harpacticoids are the prevailing taxa. These organisms can firmly attach themselves to macrophytes and find shelter and food on algal thalli. The third in abundance are amphipods (Fig. 4). They are active migrants which can move from one biotope to the other. Their migration patterns may be seasonal, but also daily. The aggregation of amphipods occasionally found in some places is likely the result of these migrations.

Examination of the interannual average estimates obtained for benthos during 1991-1992 showed low numbers of Gastropoda and a complete absence of Bivalvia on the thalli of macrophytes. There is a direct relationship between the occurrence of larvae in plankton and the recruitment of the benthos in the following year, which is especially evident in the bay.

In 1990, larvae of Gastropoda and Bivalvia were rarely found in Kruglaya bay. In the summer and autumn of 1990, the mortality rate of meroplankton increased with 2-3 times in comparison with spring. This is in accordance with the evidence obtained in

1991 were a negligible numbers of gastropods and no bivalve molluscs on *Cystoseira* were found. Correspondingly, phytophilous benthos rapidly responded to environmental changes in the bay.

Long-term variations

The tendency towards structural changes of the phytophilous benthos in Kruglaya bay is also evident through comparison between our data and those reported by E.B. Mackaveeva (1979). During that time the abundance of the main taxa of phytophilous benthos considerably decreased and the percent ratio between the relevant taxonomic groups changed (Fig. 5).

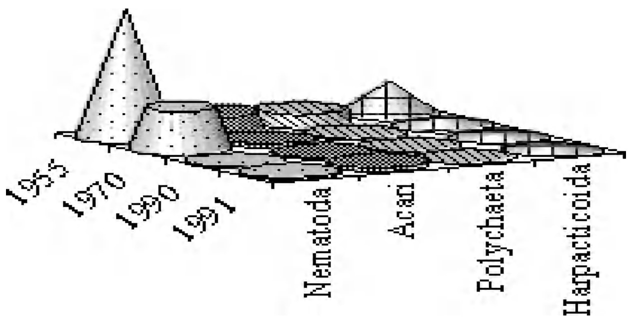


Fig. 5. Long-term variations in abundance (ind.kg^{-1} algal wet weight) for the major meiobenthos taxa in the Kruglaya Bay.

The database and seasonal index

We propose the seasonal index introduced by Nemchinov (1967) to describe interseasonal dynamics of abundance. The index is calculated based on the method of describing time series using growth rate estimates. If the index is calculated from biomass (abundance) estimates, then chain growth rates, *i.e.* ratio between the biomass (abundance) estimates of the corresponding trophic groups of the community should be determined:

$\tau_1 = y_2 / y_1$; $\tau_i = y_{i+1} // y_i$; $\ldots, \tau_{n-1} = y_n // y_{n-1}$

where y_1, $y_2, \ldots, y_i, \ldots, y_n$ are the biomass (abundance) estimates of a trophic group of the community for each following time span (month, season, etc.). The first estimate in the studied series, *e.g.* chain ratio between the biomass of seston-eaters in February and in January, is assumed to be equal to 1. Using the method of chain products, a transformed average is calculated by the formula:

$\tau_{1\,tr.}=1; \tau_2^{tr.}=\tau_2*1; \tau_3^{tr.}=\tau_3*\tau_2^{tr.};...;\tau_I^{tr.};... \tau_n^{tr.}=\tau_n*\tau_{n-1}^{tr.}$

where $\tau_1, \tau_2,..., \tau_i,..., \tau_n$ are the present growth rate, and $\tau_1^{tr.}, \tau_2^{tr.},..., \tau_I^{tr.},..., \tau_n^{tr.}$ are the transformed growth rate.

Then annual prime average (a. p. a.) is calculated from the sum of the chain transformed estimates:

$(\tau_1^{tr} + \tau_2^{tr} +...+ \tau_{12}^{tr})/ 12 = \tau_{a.\,p.\,a.}^{tr}$

Seasonal average is determined as a prime average calculated from the chain average estimates obtained throughout the season. Average for the season (spring) preceding the observation is assumed to be equal to 1. Then averages obtained for the seasons of observation are transformed by the method of chain products.

Seasonal index (J) is determined as the ratio of the transformed (tr.) values to their annual average:

$J = (\tau_i^{tr.}* 100)/ \tau_{a.\,p.\,a.}^{tr.}$

When computing the seasonal index of the meiobenthos, annual average abundances of the examined structural components of the community were used as a basis.
The seasonal index computations for the main taxonomic groups of meiobenthos in the algal communities, calculated from the number of estimates obtained in different sites during our monthly monitoring, obtained a variable seasonal index indicating different environmental and ecological conditions of the studied localities. The estimates were arranged within the range from 100 to 1000, which are the precisely determined limits in which the index may vary. Values of the index which vary from 100 to 300 can be regarded as normal seasonal cycles of the four meiobenthic taxa. The range from 300 to 1000 is characteristic for seasonal cycle fluctuations. Factors which induce these fluctuations are natural cycles of reproduction and environmental factors, such as surf force and anthropogenic load. In monitoring studies an important factor is the registered share (%) of fluctuations (seasonal index estimates greater than 300) in the seasonal cycle of taxa under study. For the open coastal sea water this makes up 26%, near the mouth of the bay 22% and inside the bay 48%. These values show how the meiobenthos on algae responds to anthropogenic load inside the bay. Among the studied taxa the percentage of fluctuations is distributed as follows: harpacticoids 32%, Acarina 26%, amphipods 21% and nematods 21%.

Conclusions

The database proved to be a useful tool in the quantitative description of spatial, seasonal, interannual and long-term variability of meiobenthos communities.

The diversity of phytophilous benthos found on the same macrophyte host, *Cystoseira crinita*, differs widely depending on the local conditions.

The analysis of seasonal dynamics of the main taxa, based on the trends in abundance and percentage peak estimates, pointed out that the seasonal modification in the structure of the benthos, associated to *Cystoseira crinita* in the Kruglaya bay, should be regarded as steady, with a constant quantitative ratio between the constituent elements. In the localities under study, the percentage of the main taxa (Acarina, Harpacticoida, Amphipoda) remains invariable.

Abundance estimates obtained for meiobenthos on algae during 1991-1992 showed a tendency of decrease, in particular for Harpacticoida and Acarina from the open coastal sea water, and for Nematoda also in the vicinity of the mouth of the bay, and for Amphipoda at each of the localities. These changes may be the result of environmental changes which took place during 1990-991; 1990 was especially unfavorable for the ecosystem of Kruglaya bay.

Long-term observation has also shown a substantial decrease in numbers of all main taxa of the phytophilous benthos and revealed changes in the structure of zoobenthos associated to the algae *Cystoseira crinita.*

Anthropogenic pollution of the coastal sea water has provoked replacement of the phytocenosis. In the Black Sea, the communities of brown algae (*Cystoseira*) are replaced by communities of green algae (*Ulveta*). This phenomenon may also involve changes in the diversity of benthic species associated to algae. The diversity of phytophilous benthos may undergo considerable changes under the replacement of one macrophyte substrate by the other.

In handling the two-year monitoring data, a seasonal index is used to identify changes in the characteristic seasonal cycle, which could identify a response to anthropogenic load in the Kruglaya bay, which is a recreation zone. This approach also allows clarifying which of the taxa occurring in the algal communities are the most vulnerable to an unfavorable environment or, on the contrary, are prosperous: which were harpacticoids and Acarina, respectively. These taxa may be used as indicators in further monitoring investigations including those focused on meiobenthos.

References

Mackaveeva H.B. 1979. Invertebrates inhabiting algae growth in the Black sea. Naukova dumka, Kiev. 228p.

Nemchinov V.S. 1967. Selected works. V.2 Science, Moscow. 498p.

Building a Global Plankton Database: Eight years after Hamburg 1996

Todd D. O'Brien

Marine Ecosystems Division – National Marine Fisheries Services - F/ST7
1315 East-West Highway – SSMC III, Silver Spring, Maryland 21044 - USA
E-mail: Todd.OBrien@noaa.gov

Abstract

The "International Workshop on Oceanographic Biological and Chemical Data Management" held in Hamburg (Germany), 1996, produced a listing of suggested metadata for plankton data management. In the eight years that followed this meeting, the efforts and experiences of adding plankton tows to the World Ocean Database profile database made it clear that there was more to building a useable plankton database than just putting plankton data and metadata into a database.

Keywords: NMFS-COPEPOD; Plankton Database; Zooplankton data; Phytoplankton data; Abundance data; Biomass data; Composition data; Quality control.

Introduction

At the "International Workshop on Oceanographic Biological and Chemical Data Management" (Hamburg - Germany, 1996), Linda Stathoplos and Todd O'Brien presented their initial efforts to include plankton tows in the World Ocean Database's profile-based architecture. The Workshop produced a listing of suggested "metadata", ancillary information about the data collection and processing methods, which should be co-stored with the plankton data to ensure usability. For eight years this metadata listing was used as a general guideline for what ancillary sampling information to store in the World Ocean Database, but it became obvious that there was more to building a useable plankton database than just putting plankton data and metadata into a database.

World Ocean Database 1998

Plankton data continued to be added to the database for two years after the 1996 workshop. Linda Stathoplos left for private industry, and the author took over leadership of the effort. In 1998, this global collection of plankton data first became public with the release of World Ocean Database 1998 (WOD98, Fig. 1, Conkright *et al.*, 1998).

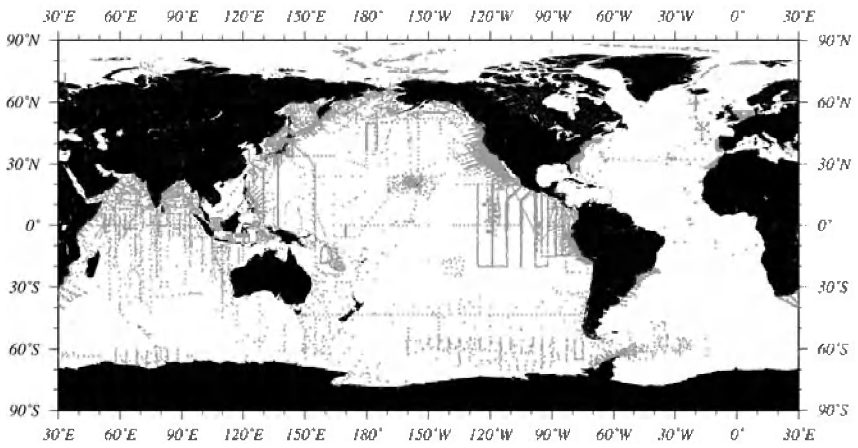


Fig. 1. Plankton tows present in World Ocean Database 1998.

While the WOD98 plankton data followed the Hamburg 1996 metadata guidelines, they were nearly impossible to use because of the complexity of the biological data itself, coupled with the complexity of the World Ocean Database profile-based data format. The plankton content of WOD98 was in a raw and basic form. The plankton species were stored without any taxonomic grouping or supplemental indexing. To extract "all copepod data", the user would have to search each and every record for the presence of any one of the over 400 unique copepod taxa contained in the database. The biomass and abundance values were also in raw form, with no quality control, stored in their original as-provided units. To utilize these values, the user would have to use sampling information and metadata to calculate a common unit. Very few users discovered these challenges, however, as the WOD98 data format, designed to efficiently manage five million temperature profiles, made finding and extracting the plankton data nearly impossible. Frustrated users frequently contacted the author directly for help.

World Ocean Database 2001

Over the next three years, more plankton data were added and the short-comings of WOD98 were addressed with the release of World Ocean Database 2001 (WOD01, Fig. 2, O'Brien *et al.*, 2002a).

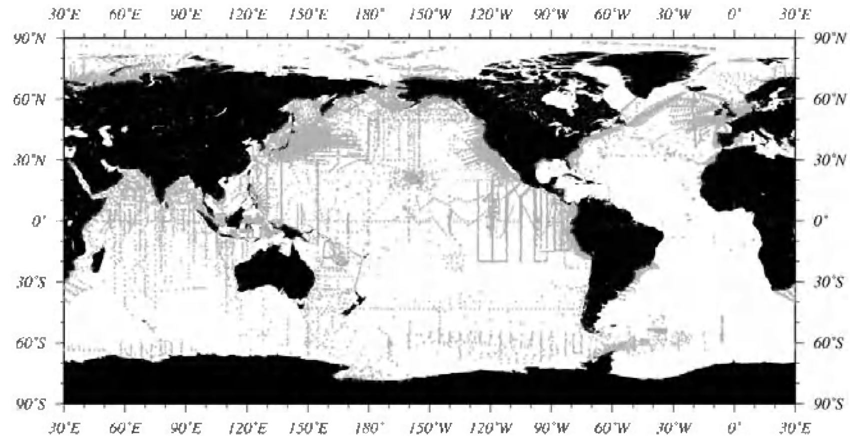


Fig. 2. Plankton tows present in World Ocean Database 2001.

To address taxonomic grouping and indexing, each plankton species was assigned a corresponding Biological Grouping Code (BGC) identification. The BGC worked as a supplemental index which would allow the user to quickly identify and access plankton by major and minor groups (*e.g.*, "zooplankton", "copepods", "chaetognaths", "phytoplankton", "diatoms", "bacterioplankton"). Common units were introduced with the addition of a Common Base-unit Value (CBV) field. Calculated from the sampling metadata, the CBV provided zooplankton in common units of "per cubic meter", and phytoplankton in units of "per liter". Basic quality control was also introduced. Mesh sizes, mouth areas, and towing depths were checked for impossible values. With the addition of the BGC and CBV, it was also now possible to access and examine all values of "phytoplankton" or "copepods" or "diatoms" with automated group-based range checking (*e.g.*, "Is this a reasonable diatom count?", "Is this a reasonable copepod count?").

While WOD01 still used the same data format and layout as WOD98, and thus had the same plankton access problems, the author provided a supplemental online plankton product called World Ocean Database Plankton (WODP). WODP was tailored to the plankton data user, offering additional documentation, content summary graphics, and plankton-specific data files and access software. While this greatly improved access to the plankton data, the access and content were still static. Searching for specific content was still not possible, and the content itself would only be updated every 3-4 years (*e.g.* WOD98, WOD01).

World Ocean Atlas 2001 - Plankton

Shortly after the release of WOD01, global mean fields of zooplankton biomass were created as part of the World Ocean Atlas 2001 series (WOA01, Fig. 3, O'Brien *et al.*, 2002b).

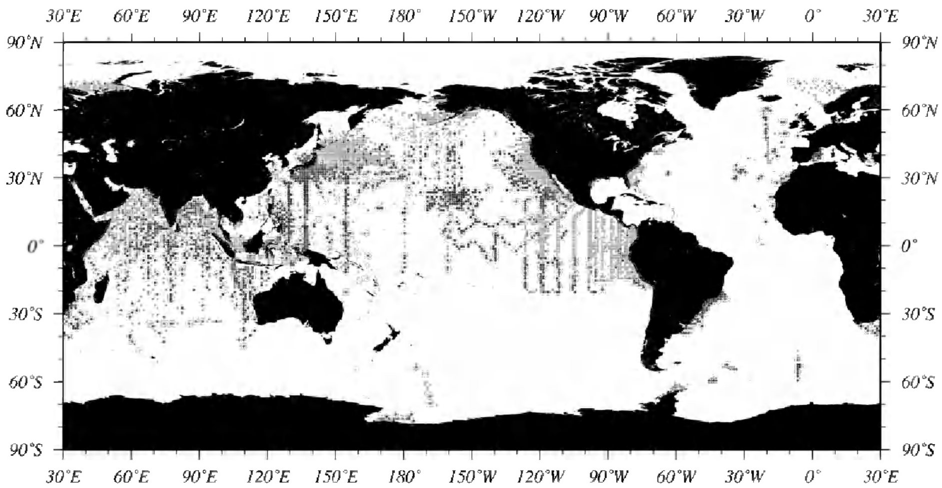


Fig. 3. Mean zooplankton biomass values present in World Ocean Atlas 2001.

During the creation of the WOA01 plankton fields, it became evident that there were still significant data access and usability issues. The creation of these fields represented the first focused effort at actually comparing and combining thousands of plankton measurements from different sampling methods and gear types. The experience not only highlighted the necessity of having complete metadata, but it also demonstrated the necessity of fully understanding and correctly translating the original plankton data and their meaning into any database.

One of the largest metadata problems involved originator-provided taxonomic group sub-total and totals. For example, an investigator reports the presence of "4 apples, 3 oranges, 2 bananas, 9 fruit, 7 vegetables". While obvious to a human reader, if the "9 fruit" is not clearly denoted as the total of all fruit types (*e.g.*, "total fruit"), an analysis program processing millions of fruit records may consider this a fourth fruit category and calculate a total fruit value of 18 from these data. Preventing these types of errors requires careful review of the metadata and data during and after processing and/or digitization. Fixing these errors means reprocessing and/or re-digitizing the mistranslated data.

During the re-processing of these mistranslated data, additional database integrity issues were discovered when comparing the reprocessed data to what was already in the database. These discoveries included lost tows, corrupted values, and disappearing metadata. The causes of these problems ranged from database software errors to limitations within the profile-based database architecture itself. While it was possible to patch and repair many of the problems, the background causes and limitations would always remain a threat to future data integrity. The best solution would be to redesign and rebuild the database.

NMFS-COPEPOD: A New Approach to Plankton Data Management

The Coastal & Oceanic Plankton Ecology, Production & Observation Database (COPEPOD) is a new effort by the National Marine Fisheries Service (NMFS) to provide quality plankton data to the research community. NMFS-COPEPOD is an online database designed specifically for plankton data, developed using the author's 10 years of hands-on experience with plankton data management. In addition to providing a complete re-processing and access to the author's previous content (*e.g.*, WOD98, WOD01), it represents a new focus on user-friendly interfaces, searching, and plankton-specific export formats. Another main focus of NMFS-COPEPOD is to provide clear credit to the associated investigators, projects, and institutes responsible for each and every data set.

NMFS-COPEPOD (http://www.st.nmfs.gov/plankton/) has been online since August 2004. New data are added and available online each month (versus every 3-4 years). As of January 2005, NMFS-COPEPOD contained 86 online data sets, with an additional 40 data sets in final processing and review. Coming in late 2005, new biomass and abundance fields will also be released (Fig. 4, O'Brien, 2005). These will be made available online and in the form of a digital atlas and database product.

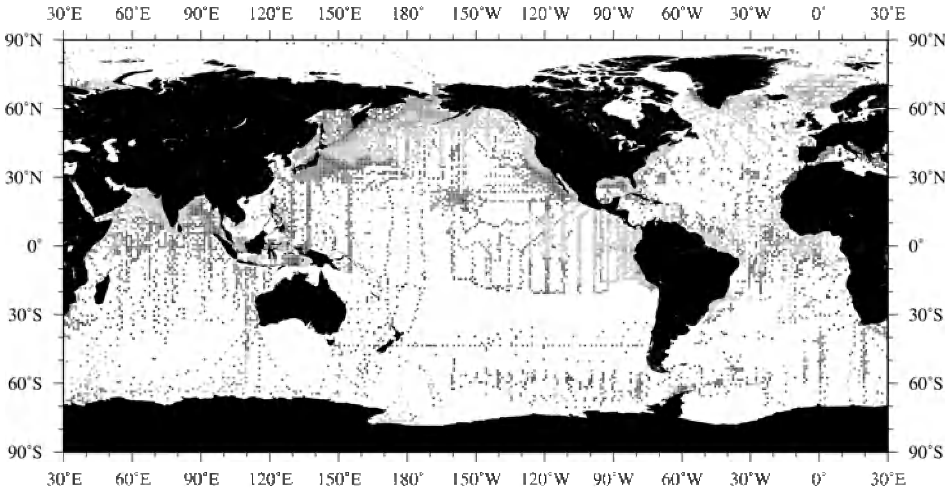


Fig. 4. Mean zooplankton biomass values coming soon in NMFS-COPEPOD 2005.

Conclusions

In the eight years since Hamburg 1996, the metadata requirements for managing plankton data have changed very little. It is the ability to apply these metadata and review the quality of the data that is necessary for managing such data. There is more to building a <u>useable</u> plankton database than just putting plankton data into a database. A plankton data manager needs to know the data (and its quirks and challenges), use the data (applying it and experiencing its short-comings and quality issues), serve the data and the needs of its users (as the reason for building it is ultimately for their use), and acknowledge the investigators (without whom there would not be any data to manage). A successfully useable plankton database should protect the quality and integrity of its data, serve its community, and thereby encourage submission of future data to the effort.

Acknowledgements

NMFS-COPEPOD is an ongoing effort by the Marine Ecosystems Division of the National Marine Fisheries Service (NMFS) Office of Science & Technology. The content of NMFS-COPEPOD is possible through the efforts and contributions of plankton scientists through the world. Too numerous to list here, the names of associated investigators, projects, and institutes are acknowledged in the "Hall of Fame" section of the NMFS-COPEPOD web site (http://www.st.nmfs.gov/plankton/).

References

Conkright M.E., T.D. O'Brien, L. Stathoplos, C. Stephens, T.P. Boyer, D. Johnson, S. Levitus, R. Gelfeld. 1998. NOAA Atlas NESDIS 25 - World Ocean Database 1998, Volume 8: Temporal Distribution of Station Data Chlorophyll and Plankton Profiles, United States Government Printing Office, Washington, DC. 129p.

O'Brien T.D. 2005. COPEPOD: A Global Plankton Database. U.S. Department of Commerce, NOAA Technical Memorandum NMFS-F/SPO-73, 136p.

O'Brien T.D., M.E. Conkright, T.P. Boyer, J.I. Antonov, O.K. Baranova, H.E. Garcia, R. Gelfeld, D. Johnson, R.A. Locarnini, P.P. Murphy, I. Smolyar, C. Stephens. 2002a. NOAA Atlas NESDIS 48 – World Ocean Database 2001, Volume 7: Temporal Distribution of Chlorophyll and Plankton Data. United States Government Printing Office, Washington, DC. 219p.

O'Brien T.D., M.E. Conkright, T.P. Boyer, C. Stephens, J.I. Antonov, R.A. Locarnini, H.E. Garcia. 2002b. NOAA Atlas NESDIS 53 - World Ocean Atlas 2001, Volume 5: Plankton. United States Government Printing Office. Washington, DC. 89p.

New high-throughput biotechnologies for sampling the microbial ecological diversity of the oceans: the informatics challenge

Carmen Palacios[1,2], Bertil Olsson[3], Philippe Lebaron[2], Mitchell L. Sogin[3]

[1]Max Planck Institute for Marine Microbiology, Celsiusstr. 1
28359 Bremen, German

[2]Observatoire Océanologique Banyuls, Laboratoire de Microbiologie
66651 Banyuls-sur-mer, France
E-mail: carmen.palacios@obs-banyuls.fr

[3]Marine Biological Laboratory. 7 MBL St.
Woods Hole, 02543 MA, USA

Abstract

Microorganisms account for most of Earth's biodiversity. They mediate key biogeochemical processes and serve pivotal functional roles in complex ecosystems, yet little is known about mechanisms responsible for the formation of microbial ecological diversity patterns. High-throughput molecular biology provides a powerful tool for measuring and monitoring patterns of microbial diversity. SARST-V6 (Serial Analysis of V6 Ribosomal Sequence Tags) is a promising technology that uses short DNA sequence tags to fingerprint the composition of microbial communities. To efficiently interpret the large amount of diversity information generated by this high-throughput technique we have made significant improvements to our SARST-V6 data acquisition and analysis informatics tools, which is now available through the WEB portal http://www.obs-banyuls.fr/UMR7621/SARST-V6.

Keywords: High-throughput microbial community analysis; Microbial ecological diversity; SARST-V6; Sunken woods.

Background information

Understanding patterns of variation in microbial populations is of great importance because these relatively simple organisms account for the majority of biodiversity on earth where they mediate key processes that sustain all forms of life. For instance, microorganisms may represent as much as 90% of biomass in marine systems where they serve key roles in remineralization of carbon, with and without oxygen, nitrogen cycling, and biogeochemical transformations of sulfur, iron and manganese (Kirchman, 2000). Microbial ecological diversity investigations in concert with detailed descriptions of environmental parameters promise to unveil new insights about interactions between microorganisms and their habitats (Green *et al.*, 2004; Horner-Devine *et al.*, 2004). Early studies of microbial molecular diversity relied upon sequence analyses of ribosomal RNAs (rRNAs) or their coding regions (Hugentzholtz *et al.*, 1998). Although rich in information, these investigations are expensive to perform and usually provide no reliable abundance data of the different kinds of organisms at a study site because of its limited throughput. To address this problem, investigators turned their attention to relatively rapid profiling methods such as terminal restriction fragment length polymorphisms (T-RFLP, Moeseneder *et al.*, 1999) or denaturing gradient gel electrophoresis (DGGE, Muyzer *et al.*, 1993). These techniques provide estimates of the relative number of specific rRNA amplicons generated by polymerase chain reaction (PCR) experiments but without accompanying DNA sequence analyses, they lack the ability to identify specific phylotypes in a given community. Moreover, since these methods simply measure relative amounts of nucleic acid from different organisms in a sample, fluorescence in-situ hybridization (FISH, Amann *et al.*, 1995) remains the most reliable method to determine the number of probed organisms. However, FISH is not a high-throughput method and it only detects organisms with rRNAs that hybridize with the designed probes.

New methods with the potential to overcome these technical difficulties are at various stages of development (Bertilsson *et al.*, 2002; Neufeld *et al.*, 2004; Kysela *et al.*, 2005). They exploit the intrinsic phylogenetic information contained in relatively short (30-150 base pairs), genetically hypervariable regions of the ribosomal RNA molecule to extract phylotype information directly from sequencing. The advantage of these technologies for ecological diversity studies is that they allow detection of all organisms present in natural samples through high-throughput sequencing while at the same time providing estimates of their relative numbers. Thus they avoid the difficulties and assumptions associated with estimating relative abundances of organisms based upon integration of band intensities generated by fingerprinting methods like DGGE or TRFLP. The high throughput generation of short sequence tags for studies of microbial diversity requires the capability to process large amounts of sequence data. In this communication we outline improvements to our informatics treatment of data from Serial Analysis of V6 Ribosomal Sequence Tags methodology (SARST-V6, Kysela *et al.*, 2005).

The informatics challenge: SARST-V6 pipeline

SARST-V6 is a molecular method that draws upon information-rich DNA sequence analysis of the 16S rRNA, while providing higher throughput and efficiency than standard small subunit ribosomal DNA sequencing protocols. The technique is modelled after serial analysis of gene expression (SAGE), which describes relative expression levels for genomic tags in mRNA populations (Velculescu *et al.*, 1995). SARST-V6 produces sequences of large concatemers of PCR-amplified ribosomal sequence tags (RSTs) from homologous V6 hypervariable regions (Kysela *et al.*, 2005). This strategy increases by at least 6-fold the yield of information about different PCR amplicons in a single sequence relative to the traditional sequencing of a single rRNA amplicon in each reaction. To extract biodiversity information from the concatemer sequences, it is necessary to identify the boundaries of each RST. Comparison against a comprehensive rRNA gene database identifies the taxonomic assignment of individual RSTs.

The flow chart in Figure 1 outlines the SARST-V6 pipeline. A pipeline consists of several scripts and programs that carry out a series of bioinformatics steps required to process data. Our pipeline aims to extract ecological diversity information from SARST-V6 data analysis. In the flow chart, customized scripts to process SARST-V6 concatemer sequences into individual RSTs are intermingled with available software (usually freeware) to analyze sequence data. All scripts particular to SARST-V6 contributed by this communication are available upon request.

From chromatogram files to sequence FASTA files

The first step in the SARST-V6 pipeline (Fig. 1) is to convert chromatograms into PHD format sequencing files using PHRED (freeware available at http://www.phrap.com/background.htm). PHD files contain not only the base pair sequence information but also the quality of each called base. To automatically trim PHD files from vector sequence and low quality reads, we use LUCY (freeware at http://www.tigr.org) with default parameters with the exception that minimum sequence length is set to 20 in order to capture single tag sequences. PHD files are converted into regular FASTA format using PHD2FASTA (http://www.phrap.com/background.htm).

From concatemer sequences to ribosomal sequence tags (RSTs)

The second stage of the pipeline identifies the boundaries of individual tags and parses the concatemer into RSTs (Fig. 1). This script recognizes imperfect punctuations that arise because of sequencing errors or failure of the type II restriction enzymes to accurately cut at their predicted cleavage sites. In addition to imperfect punctuations in SARST concatemers, there can be other artifacts generated during DNA ligation or recombinant cloning. As part of this process the software generates a SARST file (Fig. 1) that contains all RSTs and marks those that reside at the beginning or end of the concatemer as well as artefactual and truncated tags. The SARST file provides a basis for making quality control decisions about the identity and integrity of RSTs (see below). The script to generate SARST files can be run interactively over the Internet for a single concatemer sequence (see Fig. 1 for output details. URL: http://www.obs-banyuls.fr/UMR7621/SARST-V6). However, to process larger amounts of data, it will be more efficient to download the programs and scripts for use on local LINUX computers, which are publicly available in the above URL.

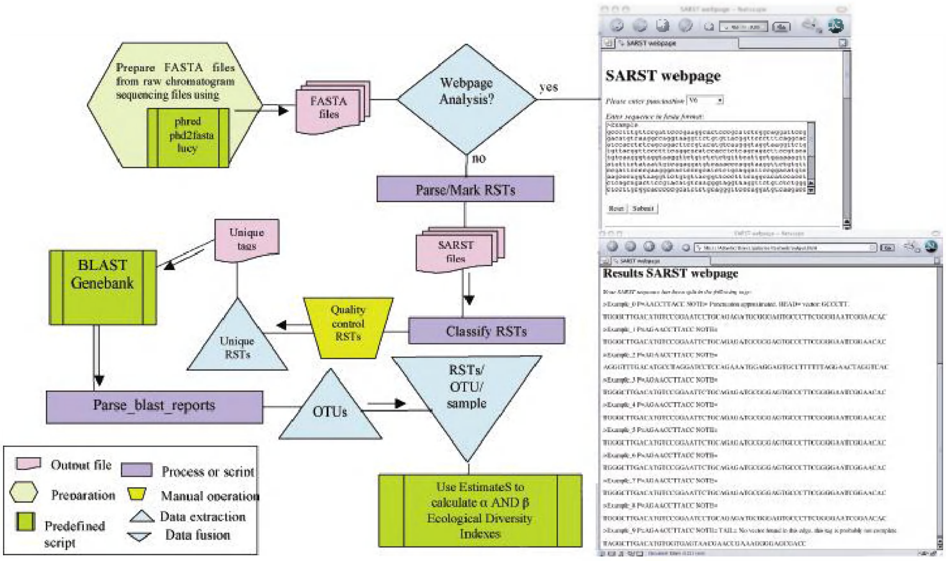


Fig. 1. SARST-V6 pipeline. This pipeline outlines the different stages of SARST-V6 sequence analysis (see text for details).

Forming integral RSTs free of artifacts

Quality control of parsed RSTs by manual inspection is accelerated through classification of RSTs according to the mark imprinted in the SARST file in previous step of the pipeline (Fig. 1). Thus we use a customized script to classify RSTs into different groups if their position in the concatemer was first, middle or last, whether or not vector boundaries are present in first and last tags and they are short, and/or if they have a particular artifact. Most of the RSTs will have no artifacts requiring no further processing. However, some RSTs in first or last position will be too short and therefore not complete or the whole SARST file lack perfectly punctuated tags. These groups of RSTs are eliminated from the dataset. RSTs that do not fall into these categories are inspected by eye for quality control. First, each classified group of RSTs according to a particular artifact is assembled into smaller, high similarity subgroups. Two programs that can assemble sequences are PHRAP (http://www.phrap.com/background.htm) and AlignIR (Technology University LI-COR, Inc). We use AlignIR 2.0.48 assemble algorithm with default parameters (minimum identity 70% and maximum successive failures 50) for this purpose. Without the assembly process, generating an alignment is not possible because of the genetic hypervariability of the region we are dealing with. Once a subgroup of RSTs is aligned, it is relatively easy to manually identify and remove the as well aligned particular artifact from all RSTs at a time.

Taxonomic affiliation of tags

After quality control of RSTs, the next step of the pipeline aims to determine their taxonomic affinity. We first pool all tags that are identical in their sequence into a unique RST (Fig. 1). A customized script will extract these unique RSTs while keeping track of how many of those tags occur in a particular environmental sample, what is necessary for estimating relative numbers of different sequence tags in the sample (see later).

The resulting unique RSTs are matched against publicly available sequence databases. We use BLAST program against nucleotide GeneBank database (http://www.ncbi.nlm.nih.gov). Resulting BLAST reports will return the organisms present in the database with the highest sequence affinity to each unique RST. We then parse these reports in a table in which unique RSTs are linked with the name of each most similar organism/s, BLAST score, e-value and sequence similarity to this organism's sequence.

Extracting OTUs and ecological analysis

Depending on the taxonomic resolution of interest (phylotype, genus, species, etc) a sequence similarity cut-off is chosen to group tags into Operational Taxonomic Units (OTUs). All tags matching a particular taxonomic group within that similarity cut-off are pooled together within an OTU.

By joining each of those extracted OTUs with the number of tags per OTU and per sample (this last value was registered in previous stage of the pipeline when extracting unique RSTs) we can then estimate species richness and evenness for each sample and β indexes of ecological diversity between samples using the freeware EstimateS (http://viceroy.eeb.uconn.edu/EstimateS).

SARST-V6 for ocean biodiversity studies: Future prospects

We have applied the SARST-V6 pipeline described in this communication to revisit the microbial diversity component of the water column and its correlation with physico-chemical parameters of the extremely acidic, high-metal laden Tinto River (Palacios *et al.*, unpublished data). Now that we have successfully explored the microbial diversity of relatively well-known environments using SARST-V6 (Kysela *et al.*, 2005; Palacios *et al.*, unpublished data), we can use this same methodology to characterize unexplored microbial communities like those that dwell sunken woods in deep waters. Sunken woods are very interesting deep-sea habitats from an evolutionary point of view as they might act as stepping-stones for chemosynthetic communities that inhabit hydrothermal vents and cold seeps (Smith *et al.*, 1989; Distel *et al.*, 2000). Our future application of SARST-V6 to sunken woods aims to explore the microbial patterns in these particular habitats and biogeochemical processes underlining them. It is likely that our results will give clues on the evolution of ocean biodiversity.

Conclusions

We have presented here our latest advances in data acquisition and analysis of SARST-V6 to illustrate the importance of Informatics when dealing with large datasets produced by high-throughput microbial community profiling methods. The SARST-V6 pipeline outlined in this communication will largely facilitate the analysis of the biodiversity generated using this sequencing technique. Computerization renders properly documented, well-organized datasets. The Ocean Biodiversity Informatics (OBI) conference statement summarizes that these characteristics allow data to be easily screened for errors, improving quality of released data. This has been our experience with SARST-V6 data analysis. Thanks to the Informatics' effort we are now ready to make publicly available our scripts and programs, hoping they will facilitate future studies of ocean biodiversity.

Acknowledgements

CP was supported by a postdoctoral stipend of the Max Planck Institute for Marine Microbiology (Bremen, Germany). We are grateful to Dave Kysela and Laura Shulman for help with the SARST-V6 script, and to Antje Boetius and Linda Amaral-Zettler for their support and encouragement.

References

Amann R.I., W. Ludwig and K.H. Schleifer. 1995. Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiololgy Reviews 59:143-169.

Bertilsson S., C.M. Cavanaugh and M.F. Polz. 2002. Sequencing-independent method to generate oligonucleotide probes targeting a variable region in bacterial 16S rRNAby PCR with detachable primers. Applied and Environmental Microbiology 68:6077-6086.

Distel D.L., A.R. Baco, E. Chuang, W. Morrill, C. Cavanaugh and C.R. Smith. 2000. Do mussels take wooden steps to deep-sea vents? Nature 403:725-726.

Green J.L., A.J. Homes, M. Westoby, I. Oliver, D. Briscoe, M. Dangerfield, M. Gillings and A.J. Beattie. 2004. Spatial scaling of microbial eukaryote diversity. Nature 432:747-750.

Horner-Devine M.C., M. Lage, J.B. Hughes and J.M. Bohannan. 2004. A taxa-area relationship for bacteria. Nature 432:750-753.

Hugenholtz P., B.M. Goebel and N.R. Pace. 1998. Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. Journal of Bacteriology 180:4765-4774.

Kirchman D.L. 2000. Microbial ecology of the oceans. John Wiley & Sons, New York. 542p.

Kysela D.T., C. Palacios and M.L. Sogin. 2005. Serial analysis of V6 ribosomal sequence tags (SARST-V6): a method for efficient, high-throughput analysis of microbial community composition. Environmental Microbiology 7:356-364.

Moeseneder, M.M., J.M. Arrieta, G. Muyzer, C. Winter and G.J. Herndl. 1999. Optimization of terminal-restriction fragment length polymorphism analysis for complex marine bacterioplankton communities and comparison with denaturing gradient gel electrophoresis. Applied and Environmental Microbiology 65:3518-3525.

Muyzer, G., E.C. de Waal and A.G. Uitterlinden. 1993. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. Applied and Environmental Microbiology 59:695-700.

Neufeld J.D., Y. Zhongtang, W. Lam and W.W. Mohn. 2004. Serial analysis of ribosomal sequence tags (SARST): a new high-throughput method for profiling complex microbial communities. Environmental Microbiology 6: 131-144.

Smith, C.R., H. Kukert, R.A Wheatcroft, P.A. Jumars and J.W. Deming. 1989. Vent fauna on whale remains. Nature 341:27-28.

Velculescu V.E., L. Zhang, B. Vogelstein and K.W. Kinzler. 1995 Serial analysis of gene expression. Science 270:484-87.

Database on Black Sea benthic diatoms (Bacillariophyta): its use for a comparative study of diversity pecularities under technogenic pollution impacts

Alexei Petrov and Elena Nevrova

Institute of Biology of the Southern Seas NAS of Ukraine
2 Nakhimov av., 99011 Sevastopol, Ukraine
E-mail: alexpet@sevinter.net

Abstract

The taxonomic database of Black Sea benthic diatom algae was created using Microsoft Office Access software and data are based on the review of available literature from 5 coastal zones: the Caucasian, Crimean, Bulgarian and Romanian coasts and the northwestern shelf (NWBS). The results of our own sampling surveys performed along the Crimean and Caucasian coasts in the period 1984-2001 were also used. The total list of Black Sea benthic diatoms holds 553 species (705 species and intraspecific taxa), pooled in 115 genera, 59 families, 31 orders, and 3 classes of Bacillariophyta. The highest species richness of diatoms is registered near Crimea and in the NWBS, representing respectively 64.2% and 69.5% of the total number of benthic diatom species ever registered in the Black Sea.

Comparative multivariate analysis of benthic diatom taxocenes from three near shore water areas of SW Crimea is done by using quantitative data on species diversity and abundance of diatoms. Those biotopes (Laspi bay, a healthy site; the open water area nearby the mouth of Sevastopol bay, a moderately polluted site and the central part of the main Sevastopol bay, a severely polluted area) differ substantially in heavy metal content (Hg, Cu, Pb, Zn, Cr and Mn) and other toxicants (DDT, PCBs, oil hydrocarbons) in the upper sediment layer (1-4 cm). Based on PCA analysis, two principal environmental components (PCs) revealed that in, through technogenically impacted locations, PC1 (55% of the total variance) is associated with the concentration gradient of several heavy metals (Pb, Cu, Mn and Cr), whereas PC2 (24%) can be associated with changes in DDT and PCB content in the upper sediment layer.

In each of the investigated areas, the specific taxocenotic diatom complexes could be statistically separated based on the results from clustering and MDS ordination, using complete linking of the Bray-Curtis similarity. The most important indicator species, which are principally responsible for the similarity within each of assemblages and the most significant discriminating species were also determined. It is proposed to consider *Tabularia tabulata, Amphora proteus* and *Navicula palpebralis* as indicators of conditionally unpolluted biotopes (Laspi bay), whereas *Tryblionella punctata, Nitzschia sigma, Caloneis liber* and *Melosira moniliformis* can be considered as indicators of water areas subject to severe technogenic impact (Sevastopol bay).

Comparative analyses show that the combination of the variables depth, Pb, Mn, Cu and PCBs can have the highest impact (Spearman rank, ρ = 0.73-0.76) on structural and diversity features of a diatom taxocene subject to different extent of toxicants.

Keywords: benthic diatoms; diversity database; taxocene structure; technogenic pollution; Black Sea.

Introduction

Although microphytobenthos is the major primary link in the trophic relationships of sublittoral ecosystems, the study of this taxonomic group is insufficiently developed in biodiversity research in the Black Sea region. Among all other groups of microphytobenthos, the benthic diatom algae (Bacillariophyta) have the highest population densities and species richness. They dwell all sublittoral biotopes, from the surf zone up to depths of about 50-70 m. They play an important role in material and energy transformation, in self-purification processes, in oxygen balance of coastal water areas and they serve as a trophic basis for larval stages of many necto-benthic species and demersal fish. Benthic diatoms are closely associated with a certain biotope and are directly subjected to environmental conditions. All this allows considering them as appropriate indicators of anthropogenic impact in complex monitoring of sublittoral ecosystems.

In contrast with phytoplankton, research on Black Sea benthic diatoms was mostly performed in the western and northwestern areas, whereas the shores of Crimea and Caucasus are less investigated. As a consequence, information on the diatom's flora is almost lacking from the southern and southeastern parts of the Black Sea. Most of the publications are devoted to floristic descriptions of species compositions and seasonal dynamics of diatom algae, whereas a minor amount of references is dedicated to the study of taxocene structures and the measurement of biodiversity based on traditionally used indexes (such as Shannon (H'), Pielou (J), etc.). Nevertheless, the application of these indexes is often inexpedient for comparative analysis of historical data in large-scale spatial and temporal analyses, especially when the frequency of sampling, the number of replicates and the sample size is unknown or when quantitative data are absent and only a species list is available.

Therefore, collection and comprehensive assessment of taxonomic and biogeographic information makes it possible to expand our current knowledge on benthic diatom structures and their specific ecological characteristics. Besides, comparative analysis using quantitative data allows revealing the changes in species structures that are subject to natural and anthropogenic influences. Finally, to perform good ecological monitoring of the Black Sea, the aim should be to apply this to at least several taxonomical groups of benthos.

The objectives of this study are 1) to integrate existing, but isolated datasets on benthic diatom diversity from several coastal regions of the Black Sea into one consolidated taxon-based database 2) to assess the effect of several heavy metals, chlorine-organic compounds and oil hydrocarbons on the structure and diversity of benthic diatom taxocenes in several near-shore water areas, which substantially differ in the level of chemical pollutants in soft sediment bottoms of the Black Sea.

Material and Methods

The taxonomic database of Black Sea benthic diatoms is based on the review of literature data (Proshkina-Lavrenko, 1963; Bodeanu, 1987-1988; Guslyakov *et al.*, 1992, 1998; Temniskova-Topalova *et al.*, 1998; Guslyakov, 2003) and from our own benthic sampling surveys performed in the period 1984-2001 along the Crimean and Caucasian coasts (Nevrova *et al.*, 2003). The taxonomic database has been created in Microsoft Office Access. According to the system proposed by Round *et al.* (1990), the total updated list of species has been prepared based on available material from 5 coastal regions of the Black Sea: the Caucasian, Crimean, Bulgarian and Romanian coasts and the northwestern shelf (NWBS).

Data on bottom sediment chemistry parameters and benthic diatoms for the assessment of pollution impact upon the taxocene diversity and structure were obtained from the comprehensive ecological surveys carried out between 1994 and 2001 (Petrov and Nevrova, 2004). In this study, three nearshore water areas of SW Crimea were compared. Laspi bay (L) is located close to a marine reserve and is almost unaffected by any technogenic pollution. The water area adjacent to the mouth of Sevastopol bay (M) is characterized by a moderate pollution level. The central part of Sevastopol bay (S) is located in the industrial zone of Sevastopol port, where the average level of toxicants in silty sediments was the highest (Fig. 1).

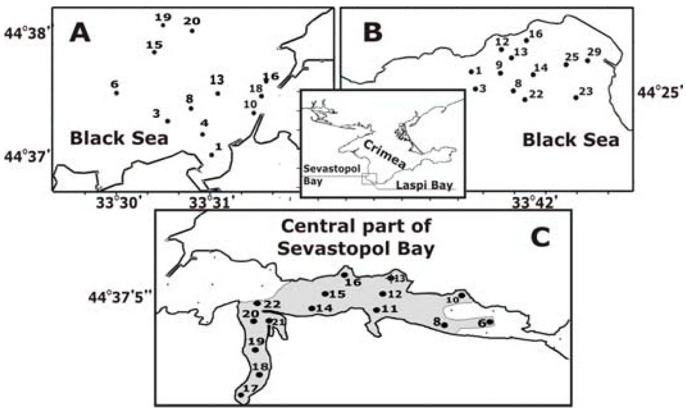


Fig. 1. Schematic map of the sampling locations in the open water area nearby the mouth of Sevastopol bay (A), in Laspy bay (B) and in the central part (grey color area) of Sevastopol bay (C).

Data on several toxicants recorded in the sediments were used to assess the effect of the pollution level on the structure and diversity features of the benthic diatom taxocene in different coastal locations. In the present study, ten toxicants were examined: 6 heavy metals, DDT, PCBs, oil hydrocarbons and bitumens. Comparison of the silty sediments of the three coastal regions showed a pronounced difference in both average values and variation range of the toxicant concentrations (Table I).

Table I. Content (average values) and the variation range (in brackets) of toxic substances recorded in the bottom sediments of the southwestern Crimea area.

Toxicant/Region		**Laspi bay**	**Mouth of Sevastopol bay**	**Central part of Sevastopol bay**
Heavy metals, mkg.g^{-1} DW	Hg	0.04 (0.03-0.05)	0.32 (0.15-0.88)	1.11 (0.40-1.88)
	Cu	7.36 (3.40-11.32)	26.17 (20.00-36.55)	8.98 (4.14-19.2)
	Pb	3.69 (3.50-5.00)	25.21 (15.0-37.5)	24.93 (9-39)
	Zn	12.0 (6.0-33.0)	18.17 (3.8-61.2)	32.69 (19.13-48.29)
	Cr	1.91 (1.51-2.62)	10.73 (7.5-20.0)	11.56 (0.53-31.79)
	Mn	6.3 (1.6-7.0)	178.3 (140.0-230.0)	6.11 (2.83-12.68)
COC, ng.g^{-1} WW	DDT	2.8 (1.8-3.0)	64.2 (14-247)	Not measured
	PCBs	5.4 (6.0-8.0)	155.0 (40-604)	1702.1 (711-3770)
CEB and oil H/C, mg.g^{-1}	CEB	0.1 (0.05-0.2)	1.3 (1.2-2.3)	1.8 (0.6-3.2)
	OHC	0.11 (0.09-0.16)	0.38 (0.14-0.90)	7.20 (1.46-15.36)

Note: References on heavy metals and chlorine-organic compounds (COC) were used: Laspi bay (Medinets *et al.*, 1994; Orlova, 1994), mouth of Sevastopol bay (Anon, 1994), Sevastopol bay (Petrov *et al.*, 2005). Data on CEB and oil hydrocarbons (Oil H/C) (Mironov *et al.*, 1992; Polikarpov *et al.*, 1992; Osadchaya *et al.*, 2003). Hg in Sevastopol bay by Kostova S.K. (unpublished data).

The sediment samples (taken from upper layer of 1-4cm) for chemical and biological analysis were collected using a Petersen grab, which was deployed on silty/sandy substrates with a depth range of 8-32m (Nevrova *et al.*, 2003). The quantitative counting of common species was performed and recalculated to 1 cm^2 of seabed. The minimum rated value of diatom abundance was assigned to 250 cells.cm^{-2} (for Sevastopol bay) and 78,600 cells.cm^{-2} (for Laspi bay and the mouth of Sevastopol bay). Species densities found in the samples, but not included in the quantitative calculation was converted to a conventional minimum value of 10 cells•cm^{-2}. A complete taxonomic analysis of diatoms on slides, prepared according to the the standard technique of cold burning in acids (Guslyakov, 2003), was carried out.

Comparative analysis of the diatom taxocene structure and diversity features has been carried out by using multivariate statistical routines included in the PRIMER package (Carr, 1997; Clarke and Warwick, 2001). Multivariate techniques, including clustering, PCA and nMDS ordination, was used to distinguish the station grouping in relation to different levels of anthropogenic pressure (Clarke, 1993; Carr, 1997). Cluster analysis was fulfilled by a hierarchical agglomerative method employing complete-linking of Bray-Curtis similarities, after log-transformation. Ordination of environmental factors, *i.e.* chemistry sediment normalized data on several heavy metals, chlorine-organic compounds (COC), oil hydrocarbons and bitumens (see Table I), was fulfilled by PCA. The significance of the differences between separated groups of stations was tested by using permutation/randomization methods (ANOSIM test).

The SIMPER routine (Clarke and Warwick, 2001) was performed to provide additional information on the species that are principally responsible for the similarity within distinguished benthic assemblages (indicator species) and for the differences between such taxocenotic complexes corresponding to each of the considered geographical locations (discriminating species). The Spearman rank correlation coefficient (ρ) was applied to detect the correlation of the combination of environmental variables, which attain the best match for the high similarities (low rank) in the biotic (abundance data) and abiotic matrices, *i.e.* to recognize a set of abiotic factors "best-explaining" the spatial differences in benthic diatom community patterns across the surveyed bottom area.

Results and discussion

Regional peculiarities in benthic diatom diversity in the Black Sea

The most recent evaluations on total species richness of Black Sea benthic diatoms resulted in 705 species and intraspecific taxa. At the Caucasian coast, 280 species and intraspecies were found, 453 at the Crimean coast, 490 in the NWBS region (without consideration of species from brackish-water estuaries and lagoons), 270 at the Bulgarian coast and 362 at the shelf of Romania. After reviewing all the diatom species dwelling in hyper saline and brackish-water lagoons, the updated list of diatoms from the NWBS includes 576 species and intraspecies (Guslyakov, 2003) and the total number of diatoms registered for the Black Sea is set to 840 species.

The highest species richness of diatoms is registered near Crimea and NWBS representing respectively 64.2% and 69.5% of the total number of Black Sea benthic diatom species. In other investigated coastal areas, this relative index was much lower (about 40%) (Fig. 2).

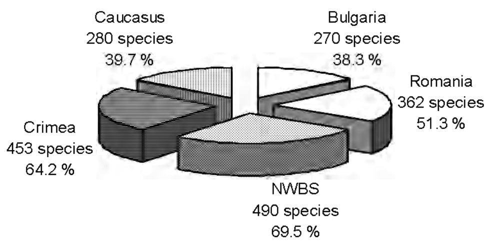


Fig. 2. Percent ratio of benthic diatom species numbers in the investigated regions of the Black Sea.

Comparing the diatom species composition from all investigated coastal regions of the Black Sea, the highest extent of species similarity occurred between the Crimean and the NWBS regions, where the Bray-Curtis similarity coefficient reached up to 77%. The lowest level of species composition similarity was found between the Crimean region of the Black Sea and the Bulgarian coast (53%) (Table II).

Table II. Cross-comparison of the diatom species composition from all investigated areas of the Black Sea, based on the Bray-Curtis similarity coefficient

	Bulgaria	Romania	NWBS	Crimea
Romania	52.5	*	*	*
NWBS	54.2	64.3	*	*
Crimea	53.1	57.2	77.0	*
Caucasus	53.8	56.2	60.5	70.4

The reported species and least inclusive taxa have subsequently been classified into five (I-V) groups, according to their frequency of occurrence (based on the reviewed literature data). In all the investigated regions of the Black Sea, 115 species and intraspecies belonging to group I (occurrence frequency 100%) were common to all regions (Fig. 3). Within this group, there are both common dominant species as well as infrequent and rare species making up the leading complex that never achieve high abundance, but which are permanently present in all the areas. Group II, III and IV (occurrence at 4, 3 and 2 regions, respectively) contain 99, 112 and 167 species and intraspecies, respectively. These are usual and ordinary numbers for benthic diatom assemblages in the Black Sea. Group V is the most numerous and being represented by 212 species, which have cited only once.

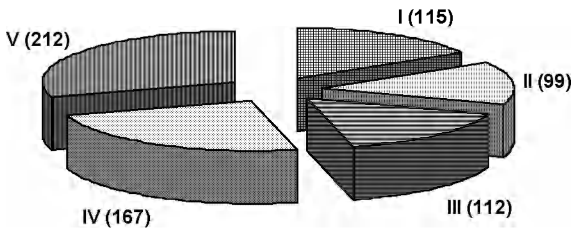


Fig. 3. Occurrence frequency of benthic diatom species of the Black Sea, divided in five groups: I – 100% occurrence; II – 80%, III – 60%, IV – 40 %, V – 20%.

According to recent data on diatom systematics, the most updated and complete list of benthic diatoms of the Black Sea includes 705 species and intraspecific taxa, 115 genera, 59 families, 31 orders, and 3 classes of Bacillariophyta. Of all the benthic diatom species observed, 76.3% are representatives of the class Bacillariophyceae, belonging to 9 orders, 30 families, 60 genera, 538 species and intraspecies. The class Coscinodiscophyceae (12.6%) is represented by 13 orders, 19 families, 28 genera, 89 species and intraspesies, the class Fragilariophyceae (11.1%) by 9 orders, 10 families, 27 genera and 78 species and intraspecies.

The following families are the most represented in the Black Sea: Bacillariaceae (6 genera; 86 species and intraspecific taxa), Catenulaceae (2; 65), Naviculaceae (3; 71), Cocconeidaceae (2; 30), Surirellaceae (4; 32), Diploneidaceae (1; 33), Cymbellaceae (4; 30) and Pleurosigmataceae (4; 28). The highest richness at the genus level was observed for the family Fragilariaceae (16 genera, 37 species and intraspecific taxa).

The most abundant species of benthic diatoms of the Black Sea, which determine the quantitative development of microphytobenthos assemblages, are *Melosira moniliformis* (O. Mull.) Ag., *Striatella delicatula* (Kutz.) Grun., *Rhabdonema adriaticum* Kutz., *Grammatophora marina* (Lyng.) Kutz., *Tabularia tabulata* (Ag.) Snoeijs, *Licmophora ehrenbergii* (Kutz.) Grun., *Achnanthes brevipes* Ag., *Cocconeis scutellum* Ehr., *Navicula pennata* A.S. var. *pontica* Mer., *Navicula ramosissima* Ag., *Berkeleya rutilans* (Trent.) Grun., *Diploneis smithii* (Breb.) Cl., *Caloneis liber* (W.Sm.) Cl., *Trachyneis aspera* (Ehr.) Cleve, *Pleurosigma angulatum* (Queck.) W.Sm., *Amphora proteus* Greg., *Amphora coffeaeformis* (Ag.) Kutz., *Bacillaria paxillifer* (O.Mull.) Hend., *Nitzschia closterium* (Ehr.) W.Sm., *Campylodiscus thuretii* Breb.

In the last 20 years, a number of species were discovered as a new to science: Achnanthes bacescui Bodeanu, Amphora karajevae Gusl., A. macarovae Gusl., Amphora lydiae Gusl., Amphora pogrebnjakovii Gusl., Amphora pontica Gusl., A. proschkiniana Gusl., A. chadjibeiensis Gusl., A. genkalii Gusl., A. topaschevskii Gusl., Cocconeis placentuloides Gusl., Cocconeis kujalnitzkensis Gusl. et Geras., Cymbella odessana Gusl., Cyclotella convexa Bodeanu, C. undulata Bodeanu, Gomphonemopsis domniciae (Gusl.) Gusl., Lyrella phyllophorae Gusl., Navicula gomphonematoides Gusl., Navicula plicata Bodeanu.

There were some new species for the entire Black Sea: *Achnanthes pseudogroenlandica* Hend., *Amphora* sp., *Hantzschia marina* (Donkin) Grunow, *Nitzschia sigmoidea* (Ehr.) W.Sm., *Undatella quadrata* (Breb.) Paddock et Sims. Some species, such as *Cocconeis britannica Naegeli, Pinnularia trevelyana (*Donk.) Rabenh., *Toxonidea insignis* Donk. and *Raphoneis amphiceros* Ehr., have not been found in the Black Sea since XIX century. Twenty-one species are rare and 48 are newly reported species for the Crimean coastal water areas (Nevrova *et al.*, 2003).

The recent increase of diatom species richness, which has been recorded in the last decades, can be a result of intensification of scientific research, but is certainly also due to more active introductions of new species into the Black Sea.

Changes in the diatom taxocene structure under different degrees of technogenic pollution

As shown above, the biodiversity of benthic diatom algae in the coastal waters around Crimea is the most investigated and attains the highest values compared with other coastal regions in the Black Sea. There are now detailed quantitative data available from the Crimean coasts that can be used as a basis for comparative analysis and assessments of changes in benthic diatom diversity patterns in relation to various environmental impacts. Considering this, results of comprehensive floristic and taxonomical surveys in several locations of SW Crimea were used not only for updating the database, but also to perform comparative analysis of diatom diversity changes in accordance with the level of anthropogenic impact.

The contamination gradient across the anthropogenically impacted water areas (M and S) has been observed through the heavy metals and chlorine-organic compounds (COC) concentrations. The possible effect of toxicants on benthic diatom diversity was assessed. Based on the results of PCA analysis, two principal environmental components (PCs) could be distinguished: PC1 (resolving 55% of total variance) is associated with the concentration gradient of several heavy metals (Pb, Cu, Mn and Cr), while PC2 (24% of the total variance) is mainly associated with changes in COC (DDT and PCBs) content in the upper sediment layer (1-4cm).

After clustering, based on the Bray-Curtis similarity index, 39 stations taken from all 3 sites, with a similarity level of about 30%, were subdivided into 3 well-distinguished groups, corresponding to 3 main sampling locations of SW Crimea (Fig. 4).

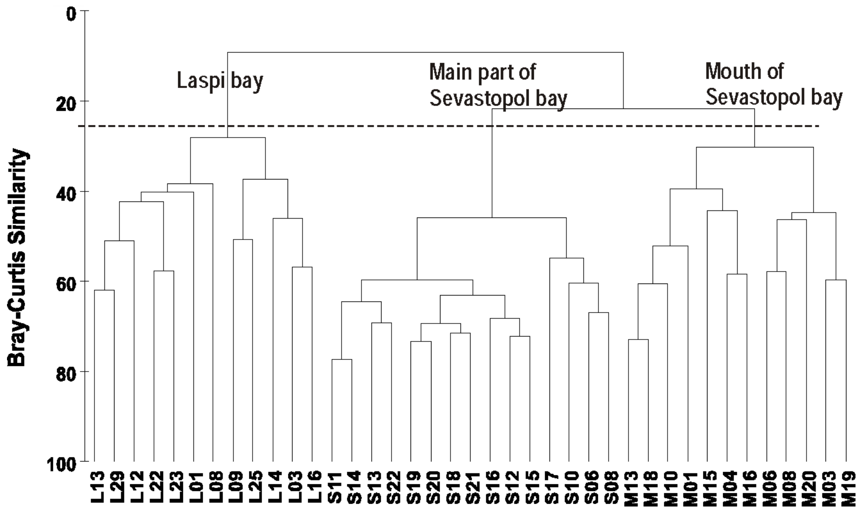


Fig. 4. Dendrogram representing the relative similarity of stations (based on Bray-Curtis similarity of log-transformed diatom abundance). The dotted line indicates the integration level (about 30%) of clusters into taxocene complexes for three coastal locations.

Results of an MDS ordination performed with the samples taken at Laspi bay (L), Sevastopol bay (S) and the open water area near the mouth of Sevastopol bay (M), show three non-overlapping areas (fig. 5).

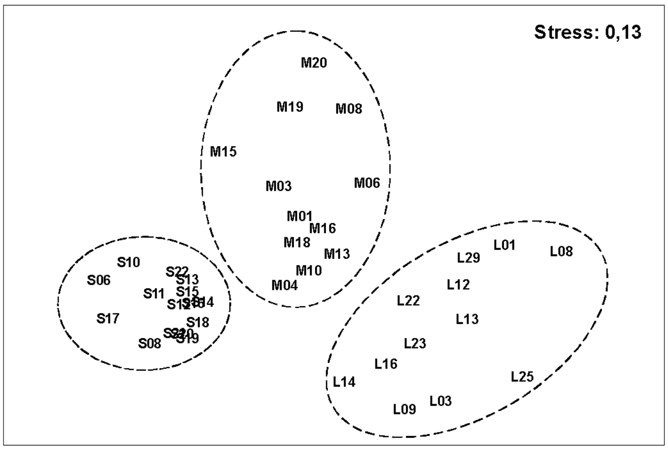


Fig. 5. The results of ordination (MDS) analysis: grouping of stations into complexes from Bray-Curtis similarity of diatom algae abundance (log-transformed). Samples are from Laspi bay (L), Sevastopol bay (S) and the mouth of Sevastopol bay (M).

Low stress function values (0.13) from the MDS analysis indicate a reliable allocation of the sample projection on 2-D plot. Differences between groups were statistically significant: the value of global R-statistics was high (0.84) at a significance level of 0.1%; pair wise testing resulted in R_p values from 0.73 to 0.94 (0.1%). These results also verify that each of the compared coastal locations is characterized by a certain taxocene complex of diatom algae. The average values of the taxocene diversity parameters for the three groups of stations are presented in Table III.

Table III. Average abundance and other species diversity parameters for 3 taxocene complexes of benthic diatoms in different coastal locations in southwestern Crimea.

Region	**Average abundance (106 cells • cm-2)**	**Total number of species**	**Number of common species**	**Number of rare species**
Laspi bay	3.020±0.562	176	53	123
Mouth of Sevastopol bay	2.572±0.413	128	38	90
Central part of Sevastopol bay	0.068±0.018	146	86	60

Differences in structure and diversity patterns between groups of stations can be explained by the influence of environmental factors (mostly technogenic impact) on the structure and quantitative development of diatom complexes. Results of SIMPER data analysis provided additional information concerning the species (indicator and discriminating ones) which are mainly responsible for the similarity within each of the distinguished taxocene complexes and for differences between such complexes (Table IV).

Table IV. Contribution from the most significant species (indicator species) into average similarity within taxocene complexes of benthic diatoms in comparing locations of southwestern Crimea

Species in comparing locations	*N*, **cells•cm^{-2}**	S_i	*S*	S_i *(%)*
L - average similarity 41.6%				
Tabularia tabulata (Ag.) Snoeijs	1139775	4.0	1.6	9.6
Amphora proteus Greg.	150667	3.5	1.9	8.5
Navicula pennata A.S. var. *pontica* Mer.	216392	2.4	1.1	5.8
Navicula palpebralis Breb. var. *semiplena* (Greg.) Cl.	98350	2.1	1.0	5.0
Grammatophora marina (Lyngb.) Kutz.	72158	1.4	1.0	3.5
Diploneis smithii (Breb.) Cl.	45867	1.2	1.1	2.8
Pleurosigma angulatum (Queck.) W.Sm.	59033	1.1	1.0	2.8
Fallacia forcipata (Grev.) Stick et Mann	32800	1.1	1.6	2.7
Cocconeis scutellum Ehr. var. *parva* Grun.	45908	1.0	0.7	2.5
Amphora coffeaeformis (Ag.) Kutz.	58975	1.0	0.9	2.4
Bacillaria paxillifera (O. Mull.) Hend.	52392	1.0	0.8	2.4
Other species				51.9
M – average similarity 43.1%				
Navicula pennata A.S. var. *pontica* Mer.	349108	4.8	1.2	11.2
Diploneis smithii (Breb.) Cl.	209275	4.1	1.3	9.5
Tryblionella punctata W. Sm.	104625	2.6	1.1	6.0
Cocconeis scutellum Ehr.	122050	2.0	1.0	4.7
Caloneis liber (W. Sm.) Cl.	226975	1.8	1.0	4.3
Nitzschia sigma (Kutz.) W. Sm.	104542	1.8	0.9	4.2
Fallacia forcipata (Grev.) Stick et Mann	157133	1.7	0.7	3.9
Ardissonea crystallina (Ag.) Grun.	104625	1.6	0.7	3.7
Other species				52.4
S – average similarity 61.2%				
Tryblionella punctata W. Sm.	8583	3.1	2.6	5.1
Diploneis smithii (Breb.) Cl.	6909	2.9	5.9	4.8
Nitzschia sigma (Kutz.) W. Sm.	3248	2.7	5.9	4.5
Caloneis liber (W. Sm.) Cl.	3719	2.7	5.4	4.4
Melosira moniliformis (O. Mull.) Ag.	2970	2.4	3.0	4.0
Tabularia gaillonii (Bory) Bukht.	1166	2.1	4.3	3.5
Navicula cancellata Donk.	2870	2.0	1.9	3.3
Grammatophora marina *(Lyngb.) Kutz.*	2235	1.9	2.6	3.1
Lyrella abrupta (Donk.) Gusl. et Kar.	1932	1.8	2.6	3.0
Cocconeis scutellum Ehr.	801	1.7	2.7	2.9
Other species				61.4%

Note: *N,* cells•cм$^{-2}$- average abundance of ***i-th*** species in taxocene complex, ***S*** – similarity function, S_i – absolute and $S_i(\%)$ – the relative contribution of ***i-th*** species in average Bray-Curtis similarity within the benthic taxocene complexes.

The average similarity of stations within each of the pollution-related taxocene complexes, evaluated by the Bray-Curtis similarity coefficient, appeared to be rather low for complex L (41.6%) and M (43.1%), whereas in complex S - where environmental conditions (pollution level) are remarkably different from the two other biotopes - average similarity value was highest (61.2%).

In the taxocene complex of the unpolluted Laspi bay, the 11 most significant indicator species (of the total list of 176) determining structural features of the taxocene, bring about 48% of total input into the average similarity within this complex. *Tabularia tabulata* and *Amphora proteus* are the top ranged species of this list (combined relative contribution is 18.1%). In the severe polluted Sevastopol bay, about 39% of the total contribution to the average similarity within the complex is determined by a group of ten top ranged indicator species (of the total list 146). *Tryblionella punctata* var. *punctata, Nitzschia sigma, Caloneis liber* and *Melosira moniliformis* are the leading taxa, displaying the highest values (5.1 – 4.0%) of their relative contribution. These parameters define the indicator role of these species in a given taxocene complex, which is formed under strong technogenic impact on the biotope. In the complex corresponding to the moderately contaminated open water area (M), a list of 8 top-ranged indicator forms (of the total list of 128) was represented by species which can be found in both polluted and healthy environments, such as: *Navicula pennata, Diploneis smithii, Cocconeis scutellum, Fallacia forcipata* and a few others.

While comparing the lists of top ranged indicator species of the two most polluted complexes (L and S), from 18 species and intraspecies, only three appeared to be common. Such a low affinity level (1/6) indicates a pronounced eco-floristic difference between these complexes, probably caused by a different tolerance of most indicator species to pollution. Besides, a high dissimilarity level (average dissimilarity amounted 72.2%) was also revealed when comparing taxocene complexes in the surveyed bays. This testifies to the significant differences between the compared water areas in species structure of a taxocene and the quantitative development of key species.

The most significant indicator species proposed by their relative contribution into average similarity within each of the complexes can also be considered as a discriminating species, hereby contributing extensively to species structure dissimilarity between taxocene complexes in the 3 compared biotopes differentiated by the degree of anthropogenic load. It is proposed to consider *Tabularia tabulata, Amphora proteus* and *Navicula palpebralis* as indicators of conditionally unpolluted biotopes (Laspi bay), whereas *Tryblionella punctata* var. *punctata, Nitzschia sigma* var. *sigma, Caloneis liber* and *Melosira moniliformis* can be considered as indicators of biotopes subject to technogenic impact.

Through a comparative evaluation, the Spearman rank correlation coefficient (ρ) showed that the combination of the variables: depth, Pb, Mn, Cu and PCBs, are best correlated (ρ = 0.73-0.76) with the alteration in structure and diversity patterns of diatom taxocene at a different extent of toxicants.

Conclusion

The database allowed revealing a contemporary state of the art on benthic diatom diversity in the coastal zone of the entire Black Sea. The total list of Black Sea benthic diatoms includes 705 species and intraspecific taxa, pooled in 115 genera, 59 families, 31 orders, 7 subclasses and 3 classes of Bacillariophyta. The highest diatom species richness is registered near Crimea and in the northwestern shelf (NWBS) representing 64.2% and 69.5% respectively of the total number of benthic diatom species in the Black Sea. The comparison of the diatom species composition from the five investigated coastal areas of the Black Sea showed that the highest level of species similarity was revealed between the Crimean and the NWBS area, where the Bray-Curtis similarity index reached up to 77%.

A comparative multivariate analysis of the benthic diatom taxocenes from the three near shore water areas of southwest Crimea showed that these biotopes differ substantially in heavy metal and other pollutant concentrations in the upper sediment layer.

The certain taxocenotic complexes were distinguished in each of the investigated water areas based on cluster analysis and MDS ordination, and significant differences in species structure between the taxocenotic complexes were also revealed. The most important indicator species, which are principally responsible for the similarity within each of the complexes, were determined as well as the most significant discriminating species. Such species can be considered as indicators of the diatom taxocene' state in a comparative assessment of biotopes in different environmental conditions. *Tabularia tabulata, Amphora proteus* and *Navicula palpebralis* are proposed as indicators of conditionally unpolluted biotopes (Laspi bay), whereas *Tryblionella punctata* var. *punctata, Nitzschia sigma* var. *sigma, Caloneis liber* and *Melosira moniliformis* can be seen as indicators of water areas subject to severe technogenic impact (Sevastopol bay).

As the result of a comparative evaluation, the Spearman rank correlation coefficient (ρ) showed that the changes in the structure and diversity features of a diatom taxocene are best associated with a combination of depth, Pb, Mn, Cu and PCBs (ρ = 0.73-0.76).

References

Anon. 1994. Ecological justification the near shore marine areas destined for allocation of marifarms along coasts of the Black Sea and Azov sea. 1994. In: Technological and scientific report on the project "Molluscs". Ivanov V.N., Morozova A.L. (Ed.). IBSS NASU, Sevastopol. 77p. (in Russian).

Bodeanu N. 1987-1988. Structure et dynamique de l'algoflore unicellulaire dans les eaux du littoral Roumain de la mer Noire. Cercetari Marine "Recherches Marines" (Institutul Roman de Cercetari Marine, Constanta). Vol. 20/21:19-251.

Carr M.R. 1997. Primer user manual. Plymouth routines in multivariate ecological research. Plymouth Marine Laboratory, Plymouth. 38p.

Clarke K.R. 1993. Non-parametric multivariate analyses of changes in community structure. Aust. J. Ecology 18: 117-143.

Clarke K.R. and R.M. Warwick. 2001. Change in marine communities: an approach to statistical analysis and interpretation. 2nd edition. PRIMER-E, Plymouth. 154p.

Guslyakov N.E. 2003. Diatom algae of Benthos of the Black Sea and contiguous aquatories. Thesis of Doctor's degree. Institute of Botany, Kiev. 36p. (in Russian).

Guslyakov N.E., E.L. Nevrova and Microphytobenthos. 1998. p.41–43; 199–215. In: Black Sea Biological Diversity. Ukraine. Vol. 7. Zaitsev Yu.P., Alexandrov B.G. (Ed.). United Nations Publications, New-York. 352p.

Guslyakov N.E., O.A. Zakordonez and V.P. Gerasimuk. 1992. Atlas of diatom algae of Benthos of the Black Sea North-West part and contiguous aquatories. Naukova Dumka, Kiev. 115p. (in Rus., abstracts in English).

Medinets V.I., A.A Kolosov and V.A. Kolosov. 1994. Toxic metals in marine environment. p.47-53. In: Investigation of the Black Sea ecosystem. Vol. 1. Medinets V.I. (Ed.). Iren-Polygraph, Odessa. 158p. (in Rus., abstracts in English).

Mironov O.G., L.N. Kiryukhina and I.A. Divavin. 1992. Sanitary and biological studies in the Black Sea. Hydrometeoizdat, St-Petersburg. 115p. (in Rus., abstracts in English).

Nevrova E.L., N.K. Revkov and A.N. Petrov. 2003. Microphytobenthos. p.270-283; 288-302; 351-162. In: Modern condition of biological diversity in near-shore zone of Crimea (the Black Sea sector). Eremeev V.N., Gaevskaya A.V. (Ed.). "Ekosi-Gidrophizika" Publ., Sevastopol. 512p. (in Rus., abstracts in English).

Orlova I.G. 1994. Chlorinated hydrocarbons in the Black Sea ecosystem. p.36-46. In: Investigation of the Black Sea ecosystem. Vol. 1. Medinets V.I. (Ed.). Iren-Polygraph, Odessa. 158p. (in Rus., abstracts in English).

Osadchaya T.S., E.I. Ovsyaniy, R. Kemp, A.S. Romanov and O.G. Ignatieva. 2003. Organic carbon and oil hydrocarbons in bottom sediments of the Sevastopol bay (the Black Sea) Marine Ecol.l J. Vol. II, 2:85-93.

Petrov A.N. and E.L. Nevrova. 2004. Comparative analysis of taxocene structures of benthic diatoms (Bacillariophyta) in regions with different level of technogenic pollution (the Black Sea, Crimea). Marine Ecol. J. Vol. III, 2:72-83. (in Rus., abstr. in English).

Petrov A.N., E.L. Nevrova E.L. and L.A. Malakhova. 2005. Multivariate analysis of benthic diatoms distribution across the space of the environmental factors gradient in Sevastopol bay (the Black Sea, Crimea). Marine Ecol. J. Vol. IV, 3: 65-77. (in Rus., abstr. in English).

Polikarpov G.G., O.G. Mironov, V.N. Egorov and G.E. Lazorenko. 1992. Molismology of the Black Sea. Naukova Dumka, Kiev. 304p. (in Russian).

Proshkina-Lavrenko A.I. 1963. Benthic diatom algae of the Black Sea. – AS USSR, Moskva-Leningrad. 243p. (in Russian).

Round F.E., R.M. Crawford and D.G. Mann. 1990. The diatoms. Biology morphology of genera. Cambridge University, Cambridge, New York, Port Chester, Melbourne, Sydney. 747p.

Temniskova-Topalova D., V. Petrova-Karadjova and Microphytobenthos. 1998. p.17-18; 70-78. In: Black Sea Biological Diversity. Bulgaria. Vol. 5. Konsulov A. (Ed.). United Nations Publications, New-York. 131p.

Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System

Tony Rees[1] and Y. Zhang[2]

[1]CSIRO Marine Research, GPO Box 1538, Hobart 7001, Australia
E-mail: Tony.Rees@csiro.au

[2]Institute of Marine and Coastal Sciences, Rutgers, The State University of New Jersey, 71 Dudley Road, New Brunswick, NJ 08901-8521, USA
E-mail: phoebe@marine.rutgers.edu

Abstract

The initial release of OBIS, the Ocean Biogeographic Information System, provided a distributed search mechanism to retrieve marine species distribution records from a range of remote data providers in real time, based on a match on species scientific name and other parameters if specified. This 'fully distributed' version 1 of OBIS was upgraded in 2004 to provide improved functionality, system response times, and metadata-level information on available data via the OBIS system, by the introduction of two new components, an 'OBIS Index' comprising a species name index and a spatial index, and a local cache of commonly queried attributes of OBIS data items, refreshed on a rolling basis from the remote data providers. The conceptual, implementation and performance aspects of these developments are described in the present paper.

Keywords: Biological information systems; Biogeography; Databases; Indexing / Spatial indexing; Distributed searching.

Introduction and OBIS version 1

OBIS, the Ocean Biogeographic Information System, is conceived as a two-, three- and ultimately four-dimensional atlas of marine species distributions based on globally distributed data holdings accessed via a central portal (Grassle and Stocks, 1999; Zhang and Grassle, 2003), and is also the designated information and data management component of the Census of Marine Life (for information on the latter, see www.coml.org/). Functionally, OBIS comprises a central portal – presently located at Rutgers University, New Jersey, and accessible via www.iobis.org – which communicates with the various remote data providers via standard web protocols (XML over HTTP), while the inevitable heterogeneity of database or file structures at the provider end is standardised using 'wrapper' or translation software which enables the portal to issue common requests to, and receive back data in a common format from, any provider connected to the system.

Version 1 of OBIS was constructed in late 2001 and went live on the Rutgers site in January 2002, using a fairly standard architecture for what is effectively a fully distributed system, as illustrated in Fig. 1.

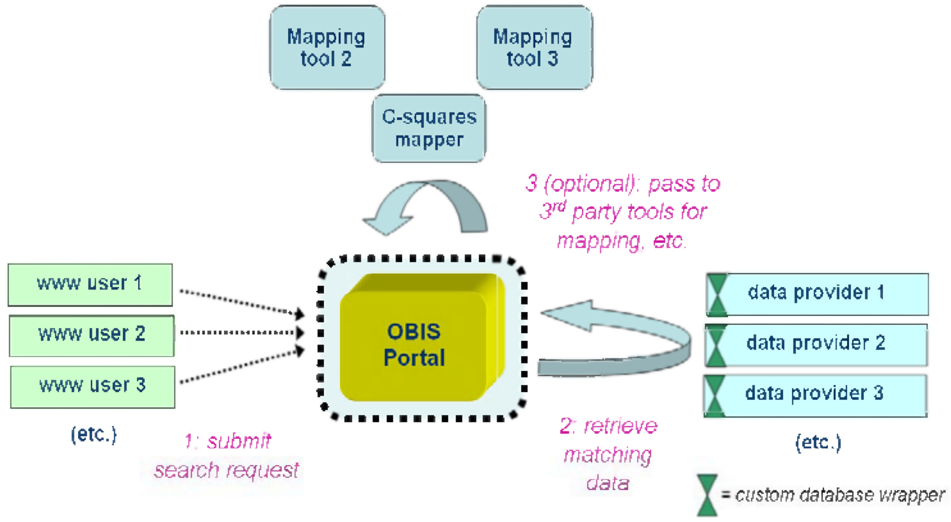


Fig. 1. Architecture of OBIS version 1, January 2002-February 2004.

In this architecture, the types of queries to be supported by the system are first designed and reflected in an XML data schema, then a custom wrapper is designed and installed at each remote data provider which will support queries on these terms and parameters (such as species scientific name, date and time of collection, water depth, and location by latitude and longitude) and return matching data to the portal upon request. Once this architecture is in place and connected via the internet, a user can submit a request for data (*e.g.* by species name or locality – labelled 1 in diagram), the portal issues the appropriate request to the remote data providers and merges any matching results returned into a single result set (2 in diagram), and the user then has further options (3 in diagram) to pass this result set to one of a number of available mapping or modelling tools as desired, or simply view or download the data to their own system for further investigation and user-specific operations.

An architecture such as this has advantages with respect to being relatively simple and rapid to design and implement, having a fairly 'light' footprint at the portal end so far as storage and maintenance are concerned, freeing the portal from any data custodianship issues, and no currency issues (information is always as up-to-date as at the remote providers). On the other hand, there are also problems with such systems which quickly become apparent in practice, and can result in a less than ideal user experience. Specific problems can be identified as follows:

- Reliability. The system is only as reliable as the 'up time' of all the contributing databases allows. If a remote data source is down, it cannot be searched.

- Speed. The system is only as fast as its slowest contributor to respond, and / or the bandwidth of the physical link to the same. Often the wait for a response may be set to a default timeout (*e.g.* 2 minutes), which means in practice that many searches take this long even if they return no data.
- User information. There is no information presented to the user in advance as to what is the coverage of the system – which species have associated data and are therefore worth searching on, and the size of the resulting dataset which will be returned (which in the OBIS case, can vary from 0 to over 40,000 records for a single species).
- Value adding. The system returns matches by scientific name, but with no added value such as an associated common name or 'organism type' (taxonomic category), and no synonym resolution since this information is not available from the remote data sources in any consistent manner, if at all.
- Serial versus parallel searching. Searches are undertaken serially, *e.g.* to discover OBIS information on the 42 known species of whales one has to undertake 42 separate searches (at up to 2 minutes per species), and searches on larger groups rapidly become impracticable (*e.g.* the 16,000+ marine fishes, or the subset of the latter beginning with 'A', etc.).
- Service chaining. In order to map (for example) the distribution of a marine species – for example *Balaenoptera physalus*, the fin whale – first, all 43,000 records must be retrieved, and only then can be sent to a mapper, *e.g.* as an XML file (which may or may not be able to cope with such a quantity of data).
- Spatial searching. Search for (for example) all data items within a given region defined as a bounding box can be slow, on account of the large quantities of data to be parsed at remote locations and returned in real time.
- Need for correct spelling. If a user enters an incorrectly spelled name, no data are returned (unless by chance a similar erroneous name exists in a contributing database), and there is no indication to the user of applicable 'near matches' owing to the nature of the query method (which searches for an exact match).

Drawing on those aspects of OBIS version 1 which had already proved successful, including building a community of remote data providers and implementing a common search protocol, planning for a 'new, improved' version 2 of OBIS started in March 2003, which would address the issues identified above with the goal of significantly upgrading usability and performance of the system.

The OBIS Index

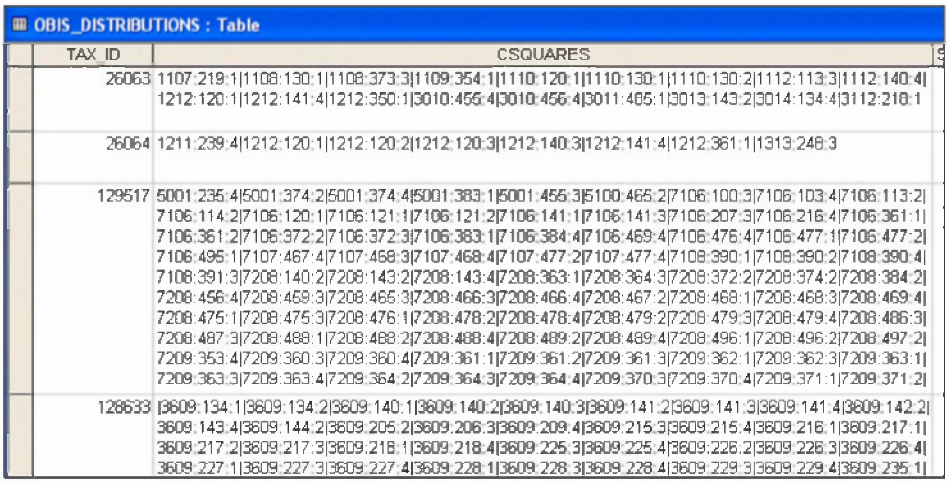
It was realised at the start of the upgrade process that incorporation of a central 'OBIS Index', to reside on the portal, would be a key concept in addressing the majority of the issues identified above, in other words, moving from a fully distributed system to a hybrid approach based on crawling the remote data providers and holding a set of summary-level information or metadata regarding each species on the portal. Such an index would then allow user searches to be split into a two stage operation: 'stage 1' searches would operate on the index and provide metadata level information on available OBIS content very rapidly, while 'stage 2' searches would retrieve actual item-level content once the user had identified exactly what data should be retrieved. In the initial prototype constructed, these 'stage 2' searches were fully distributed queries to the remote data providers as described above for OBIS version 1, while by the time the system was ready to deploy, these had been replaced by queries to a locally held data cache (see next section).

The OBIS Index was constructed to be both a name index and a spatial index. The name index holds summary information by species, such as total number of accessible records, contributing data sources, date range (earliest, latest years represented in the data set for any species), a selected common name for initial rapid display, and any synonym resolution as required, the latter drawn from recent work of the 'Catalogue of Life' project (Bisby *et al.*, 2004), together with allocation to a custom taxonomic hierarchy to support searching and grouping by taxonomic category as required, as well as presentation of an 'organism type' – examples 'a fish', 'a whale' – alongside every returned species name. An additional feature of the name index is to provide support for a 'fuzzy name matching' function via a modified version of the species name stored alongside the original. This information, together with inventory-level data such as which species names originate from which contributing databases, is stored in a small relational database which resides at the portal and provides support for 'stage 1' queries as defined above, which basically return lists of relevant species names and associated metadata in response to a user's query. Every unique scientific name available via the system is also allocated a unique (internal) numeric identifier which links the various tables together.

A supplementary component of the name index is the addition of names of species considered to be marine in the Catalogue of Life, but presently unrepresented by distribution data among OBIS' current data contributors; this allows a degree of gap analysis (assessing the percentage of known species in any particular group for which OBIS data are available over time), at least to the extent that Catalogue of Life coverage is itself complete, and also allows users to check the spelling of entered marine species names whether or not OBIS has matching species distribution data at the present time.

The spatial index forms a separate table within the 'OBIS Index' database and comprises a list of species identifiers, each associated with a set of codes which represent the spatial distribution of the available data within a set of global 0.5 x 0.5 degree squares, labelled using the 'c-squares' hierarchical notation (Rees, 2003, 2004), as shown in Fig. 2. In the current implementation, multiple c-squares codes (examples: 1107:219:1, 1108:130:1) are held as a concatenated text string using a separator character (vertical bar or 'pipe') between each code, and spatial search is by matching the code for any square entered by the user to any position in the c-squares string, in order for a 'hit' to be detected. For example, in the c-squares notation, the ten degree square extending from 10° to 20° N and 70° to 80° E is represented by the code 1107, the five degree square from 10° to 15° N and 75° to 80° E is represented by the code 1107:2, the one degree square from 11° to 12° N and 79° to 80° E is represented by the code 1107:219, and the 0.5 degree square from 11° to 11.5° N and 79° to 79.5° E is represented by the code 1107:219:1, and a search for data in any of this nested set of squares will return a 'hit' for the first species indicated (species id = 26063, which in fact corresponds to *Suggrundus macracanthus*, a species of fish). By this means, the spatial index (when cross referenced to the name index) supports queries of the type 'list all the species with data in a selected X degree square', where X belongs to the set ten, five, one or 0.5 degrees (at present, only the ten degree option is offered, in order to avoid too many queries returning no data), or optionally, a number of other spatial queries can be constructed, such as constrain by taxonomic group, etc.

Of further interest is that this example species (selected at random) is associated with 25 unique data records, but only 18 squares (at half degree resolution), in other words a degree of information compression is incorporated into the spatial index when multiple records occur in close proximity, which leads to additional efficiency for data storage and transfer (*e.g.* to relevant mappers). For example, the map shown in Fig. 6 for the minke whale, *Balaenoptera acutorostrata* (2,131 records), is generated from a list of 593 relevant squares, while that for the fin whale, *Balaenoptera physalus* (43,435 records) requires only 1,678 squares at the same resolution, a saving of over 96% in (meta-) data storage, data transmission time and required bandwidth via the web, and mapper processing time to generate the relevant 'quick map'.



OBIS_DISTRIBUTIONS : Table

<table>
<tr><th>TAX_ID</th><th>CSQUARES</th></tr>
<tr><td>26063</td><td>1107:219:1|1108:130:1|1108:373:3|1109:354:1|1110:120:1|1110:130:1|1110:130:2|1112:113:3|1112:140:4|
1212:120:1|1212:141:4|1212:350:1|3010:455:4|3010:456:4|3011:485:1|3013:143:2|3014:134:4|3112:218:1</td></tr>
<tr><td>26064</td><td>1211:239:4|1212:120:1|1212:120:2|1212:120:3|1212:140:3|1212:141:4|1212:361:1|1313:248:3</td></tr>
<tr><td>129517</td><td>5001:235:4|5001:374:2|5001:374:4|5001:383:1|5001:455:3|5100:465:2|7106:100:3|7106:103:4|7106:113:2|
7106:114:2|7106:120:1|7106:121:1|7106:121:2|7106:141:1|7106:141:3|7106:207:3|7106:216:4|7106:361:1|
7106:361:2|7106:372:2|7106:372:3|7106:383:1|7106:384:4|7106:469:4|7106:476:4|7106:477:1|7106:477:2|
7106:495:1|7107:467:4|7107:468:3|7107:468:4|7107:477:2|7107:477:4|7108:390:1|7108:390:2|7108:390:4|
7108:391:3|7208:140:2|7208:143:2|7208:143:4|7208:363:1|7208:364:3|7208:372:2|7208:374:2|7208:384:2|
7208:456:4|7208:459:3|7208:465:3|7208:466:3|7208:466:4|7208:467:2|7208:468:1|7208:468:3|7208:469:4|
7208:475:1|7208:475:3|7208:476:1|7208:478:2|7208:478:4|7208:479:2|7208:479:3|7208:479:4|7208:486:3|
7208:487:3|7208:488:1|7208:488:2|7208:488:4|7208:489:2|7208:489:4|7208:496:1|7208:496:2|7208:497:2|
7209:353:4|7209:360:3|7209:360:4|7209:361:1|7209:361:2|7209:361:3|7209:362:1|7209:362:3|7209:363:1|
7209:363:3|7209:363:4|7209:364:2|7209:364:3|7209:364:4|7209:370:3|7209:370:4|7209:371:1|7209:371:2|</td></tr>
<tr><td>128633</td><td>|3609:134:1|3609:134:2|3609:140:1|3609:140:2|3609:140:3|3609:141:2|3609:141:3|3609:141:4|3609:142:2|
3609:143:4|3609:144:2|3609:205:2|3609:206:3|3609:209:4|3609:215:3|3609:215:4|3609:216:1|3609:217:1|
3609:217:2|3609:217:3|3609:218:1|3609:218:4|3609:225:3|3609:225:4|3609:226:2|3609:226:3|3609:226:4|
3609:227:1|3609:227:3|3609:227:4|3609:228:1|3609:228:3|3609:228:4|3609:229:3|3609:229:4|3609:235:1|</td></tr>
</table>

Fig. 2. Fragment of the spatial index for OBIS version 2.

As mentioned above, the spatial index also supports the production of 'quick maps' (representation of species distributions by global 0.5 degree squares) directly from the index, in other words without requiring a (potentially slower) request to retrieve the atomic level species data for this purpose. This is achieved by rapidly assembling relevant strings of codes from the spatial index into the HTML page of search results in advance, so that the user is presented with a set of pre-configured links which, when pressed, will submit the relevant list of squares to a web based utility at CSIRO Marine Research, the c-squares mapper (see www.marine.csiro.au/csquares/about-mapper.htm) which processes the list, plots the relevant squares on to one of a range of user-selectable base maps, and returns the result as a gif image to the user's web browser, as per the example in Fig. 6.

Figs. 3-6 illustrate aspects of current OBIS 'stage 1' searches, including two initial search pages, an example search result for information on 'all whales', and an example 'quick map' for a user's selected species, all drawn from the holdings of the Index, that is, without requiring any connection to the item-level data at this point.

Fig. 3. OBIS version 2 start page, as at November 2004 – including 'click-on-a-map' spatial search, and express search input text boxes for scientific and common name searches.

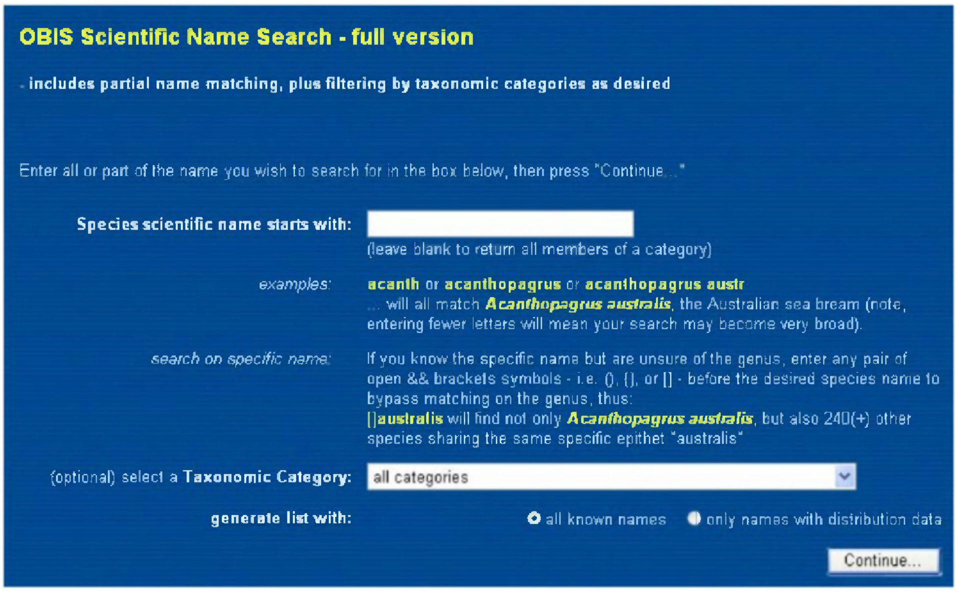
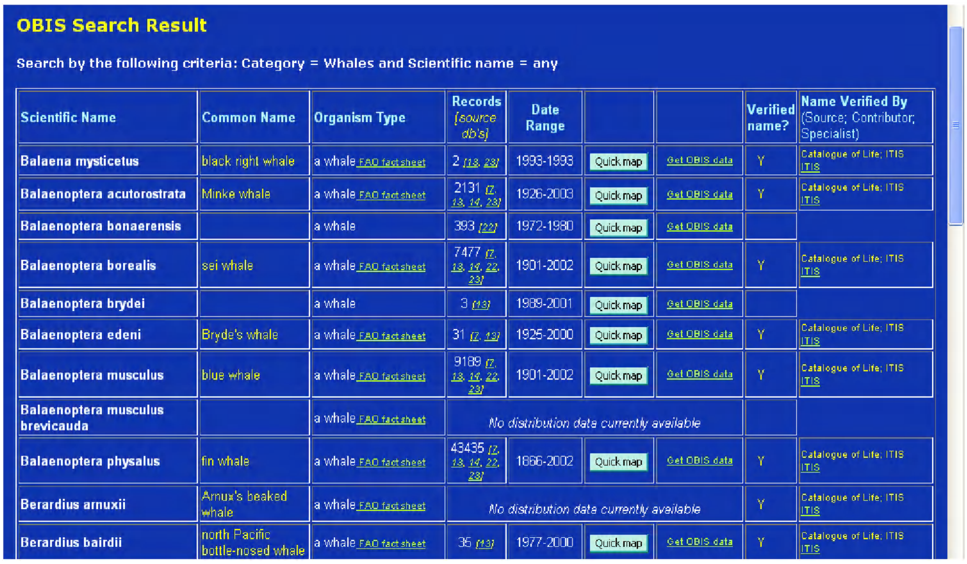


Fig. 4. OBIS version 2 full scientific name search page, as at November 2004 – including 'partial name matching, and filter by taxonomic category.



Scientific Name	Common Name	Organism Type	Records [source db's]	Date Range			Verified name?	Name Verified By (Source; Contributor; Specialist)
Balaena mysticetus	black right whale	a whale FAO fact sheet	2 [13, 23]	1993-1993	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera acutorostrata	Minke whale	a whale FAO fact sheet	2131 [7, 13, 14, 23]	1926-2003	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera bonaerensis		a whale	393 [22]	1972-1980	Quick map	Get OBIS data		
Balaenoptera borealis	sei whale	a whale FAO fact sheet	7477 [7, 13, 14, 22, 23]	1901-2002	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera brydei		a whale	3 [13]	1989-2001	Quick map	Get OBIS data		
Balaenoptera edeni	Bryde's whale	a whale FAO fact sheet	31 [7, 13]	1925-2000	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera musculus	blue whale	a whale FAO fact sheet	9189 [7, 13, 14, 22, 23]	1901-2002	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Balaenoptera musculus brevicauda		a whale FAO fact sheet	*No distribution data currently available*					
Balaenoptera physalus	fin whale	a whale FAO fact sheet	43435 [7, 13, 14, 22, 23]	1866-2002	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS
Berardius arnuxii	Arnux's beaked whale	a whale FAO fact sheet	*No distribution data currently available*				Y	Catalogue of Life; ITIS ITIS
Berardius bairdii	north Pacific bottle-nosed whale	a whale FAO fact sheet	35 [13]	1977-2000	Quick map	Get OBIS data	Y	Catalogue of Life; ITIS ITIS

Fig. 5. Example 'stage 1' search result, OBIS version 2, comprising a list of matching species names with associated metadata, plus links to 'quick maps' and 'get OBIS data' (=stage 2) searches.

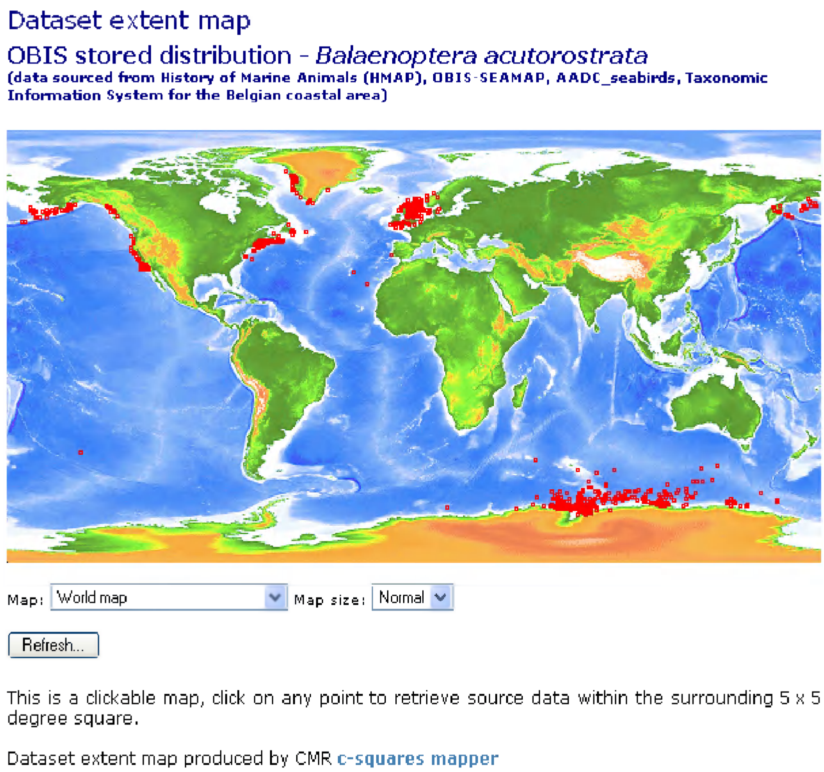


Fig. 6. Example 'quick map' generated from the spatial index holdings for a species of whale (Balaenoptera acutorostrata, 2,131 records) in a few seconds using the c-squares mapper located at CSIRO Marine Research.

'Stage 2' searches and the OBIS data cache

As described in the previous section, with the initial redesign of OBIS incorporating the new indexing functions, the requirement for 'stage 2' (= 'get data') queries is deferred in many instances until the user has familiarised his or herself with relevant system content using a 'stage 1' or index level search, leading to much faster and more satisfactory performance in the initial stages, and a reduction in the load on the system since many initial queries can be answered from the index alone (such as whether or not data exist for a species of interest, what species occur in a given area, and even the production and browsing of 'quick maps'). Nevertheless, it is essential to provide access for 'stage 2' searches from this point onwards, and in the hybrid 'index plus distributed search' architecture such queries are themselves still subject to a number of the disadvantages of a fully distributed system as described above, even with the introduction of upgraded 'wrapper' technology, introduced when OBIS moved to the DiGIR data retrieval protocol (Blum *et al.*, 2001) in place of the initial custom database wrappers, concurrently with the present upgrade.

These residual negative aspects have been addressed by the introduction of a data cache on the Portal, holding a subset of the full record (as a copy) for every OBIS data item accessible via the remote data providers, updated on a rolling basis. The purpose of this cache is to insulate the user from any individual provider being off line or unresponsive at time of querying, and also to provide a faster and more uniform response to user queries. (As a by-product, it also facilitates creation of the Index, which otherwise would require numerous and possibly slow queries to the remote providers on a species-by-species basis). Together with the Index, this cache is shown in the revised architecture as implemented for OBIS version 2, below (Fig. 7).

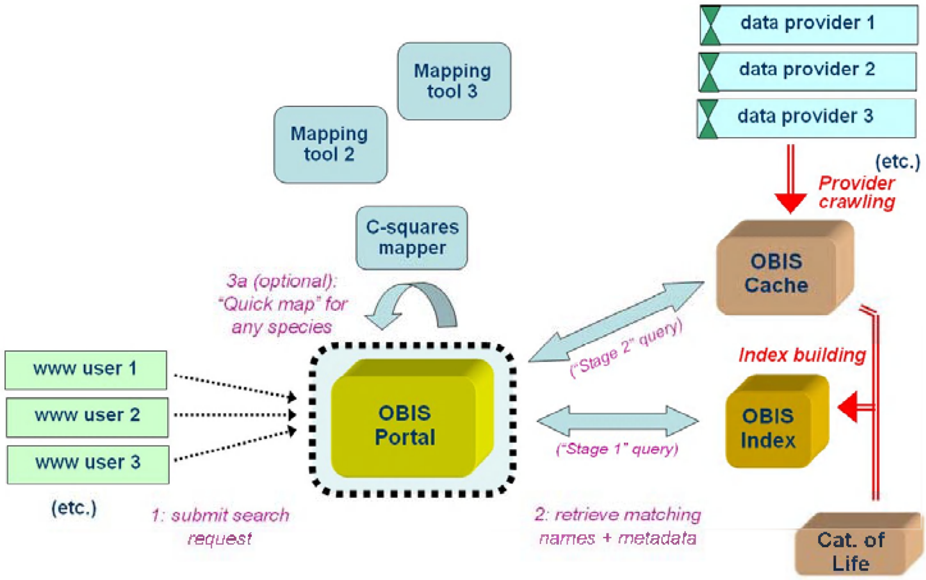


Fig. 7. Schematic of overall architecture for OBIS version 2.

The practical implementation of this architecture requires initial provider crawling to populate the cache, then creation of the index (name and spatial components) by parsing the cache content and also incorporating relevant information from the Catalogue of Life. As described above, 'stage 1' queries can then be issued against the index (relatively small in size, *e.g.* 100,000 rows at this time for the main 'obis_species' table) and used for the generation of 'get OBIS data' links and 'quick maps', while 'stage 2' queries operate against the cache (5 million + rows) but are still substantially faster (and potentially more complete) than querying the remote data sources in real time. The main disadvantages of this new architecture are its increased complexity and requirement for resources of data storage and maintenance, and the necessity to keep both cache and index content up-to-date by continual re-crawling of the remote data providers as new data are added to the system, or individual records are altered or deleted at the provider end.

Conclusion

The new version of OBIS released in 2004 has achieved a quantum leap in usability, and addressed all of the weaknesses described above for a fully distributed system, while at the same time incorporating additional innovative approaches to spatial indexing, searching and mapping, search by a custom taxonomic hierarchy, 'fuzzy' matching of species scientific names, and more. With all such systems, a degree of continuous improvement and evolution to reflect changing user demands or system possibilities will be inevitable over time, however it is felt that the current 'OBIS version 2' offers a satisfactory balance of user-weighted features as against the increased complexity and requirement for technical resources from the portal and system design points of view. Further development of OBIS will incorporate our experiences with the current system over the next 12-18 month time frame as well as the potential to exchange experiences with the developers of GBIF, the Global Biological Information Facility (www.gbif.org) and others working in similar areas of distributed biological information retrieval.

Acknowledgements

The authors are grateful for the contributions of members of the OBIS technical subcommittee (Rainer Froese, Daphne Fautin), other members of the 'OBIS web development team' (Karen Stocks, James Wood), and the OBIS International Committee (in particular Fred Grassle and Mark Costello) in reviewing progress and suggesting particular aspects for implementation. The help of Pamela Brodie and Miroslaw Ryba (CSIRO) and Lissa Jerry and Wei Zhuang (Rutgers) in assisting with practical aspects of OBIS development and deployment is also gratefully acknowledged.

References

Bisby F.A., R. Froese, M.A. Ruggiero and K.L. Wilson. 2004. Species 2000 and ITIS Catalogue of Life, Annual Checklist 2004: Indexing the world's known species. CD-ROM. Species 2000: Los Baños, Philippines.

Blum, S., D. Vieglais and P.J. Schwartz. 2001. DiGIR – Distributed Generic Information Retrieval. Powerpoint presentation, available at http://digir.sourceforge.net/events/20011106/DiGIR.ppt.

Grassle, J.F. and K.I. Stocks. 1999. A Global Ocean Biogeographic Information System (OBIS) for the Census of Marine Life. Oceanography 12(3):12-14.

Rees, T. 2003. "C-squares", a new spatial indexing system and its applicability to the description of oceanographic datasets. Oceanography 16(1):11-19.

Rees, T. 2004. Use of c-squares spatial indexing and mapping in the 2004 release of OBIS, the Ocean Biogeographic Information System. Abstract and Presentation, EOGEO 2004, London, UK. Available via the EOGEO website at http://www.eogeo.org/Workshops/EOGEO2004/.

Zhang, Y. and J.F. Grassle. 2003. A portal for the Ocean Biogeographic Information System. Oceanologica Acta 25:193-197.

Creation of the information retrieval system for collections of the marine animals (fish and invertebrates) at the Zoological Institute of the Russian Academy of Sciences

Igor S. Smirnov, Andrei L. Lobanov, Alexei A. Golikov, Elena P. Voronina and Alexei V. Neyelov

Zoological Institute
Russian Academy of Sciences 199034, St. Petersburg, Russia
E-mail: smiris@zin.ru

Abstract

The collection of the Zoological Institute RAN (ZIN) is one of the largest in the world and contains over 100,000 samples of 26,000 species of marine invertebrates and over 160,000 specimens of 8,700 species of marine and freshwater fishes and fishlike vertebrates of the world. Digitizing these catalogue data and creating a virtual library is now one of the main objectives of ZIN. The creation of the collection database on fish and invertebrates started in 1987 and is now very prospective for studying biocoenotic relationships and marine fauna ecosystems. In the course of the organization of our faunistic, ecological and collection data, the informational retrieval systems OCEAN and ECOANT (in standard ZOOCOD) were designed at the ZIN. The databases comprise Pleuronectiformes and Scorpaeniformes, as well as chiton, bivalve and brittle star collections. Different international projects are in development now, and it will be necessary to examine the experiences of these teams and from here make the attempt to create not ideal but optimal systems for the input and treatment of marine data.

Keywords: databases; zoological collection; marine fauna.

The unique collection of different animal groups, kept at the Zoological Institute RAN (over 60 million items in total), including type specimens and series, is worldwide known and is of great interest to zoological research. Today, for example, the ichthyological collection contains over 160,000 catalogued specimens (over 53,000 catalogued items) of 8,700 species of marine and freshwater fishes and fishlike vertebrates of the world. The marine invertebrate collection contains over 100,000 samples; some of them include tens and hundreds of specimens of 26,000 species. The scientific collection of ZIN is permanently supplemented and makes the number of specimens grow. The species diversity of Russian seas and adjacent waters is almost entirely represented in the collection and in large series from many localities. The participation of ZIN specialists in Russian and foreign expeditions has allowed us to obtain material from various distant areas of the world. For instance, the ZIN actively participates in research of the Southern Ocean and Antarctic biota, since the First Soviet Antarctic Expedition in 1955. Thanks to this participation a huge amount of material on the fauna of this region has been collected and for the most part catalogued (Atlas of Antarctica, 1969; Smirnov and Neyelov, 1996).

Digitizing the catalogue data and creating a virtual library has now become one of the main objectives of the largest museums of the world (Smirnov, Lobanov and Dianov, 1999). The experience of using high technology and databases in foreign countries started much earlier and was more intensive than in Russia. Today some of the largest natural history museums of the world (Natural History Museum, London; Museum National d'Histoire Naturelle, Paris; California Academy of Sciences, San Francisco; National Museum of Natural History, Washington; National Science Museum, Tokyo etc.) have web sites with electronic catalogues or collection data bases on several animal groups. Creating virtual natural history museums is promoted by these electronic catalogues and libraries, such as the electronic catalogue of invertebrates on the web site of the United States Antarctic Program http://www.nmnh.si.edu/iz/usap/usapdb.html, and FishBase: http://www.fishbase.org, a large information system with key data on all marine and freshwater fishes of the world, as well as collection data of different world museums.

The information retrieval systems and the geographic information systems not only make the work of the zoologists easier, gathering the data on species from the collection catalogue and field books and manually mapping these data, also allow us to quickly visualize the information on the occurrences of the animals from the collections that are kept at the different museums during years and centuries. This kind of software and databases will be helpful in the analysis of long-term changes of fauna composition among different regions. Together with paleontological material and geographical data about the changing of the boundaries and the position of the continents, it will allow a quick analysis of the various hypotheses on the distribution of taxa, using maps with geological reconstructions. It would be possible to retrace the history of the faunal formation and to study the influences of both climate and geological changes onto biota.

Except for the historical and cognitive value, this work has significant ecological importance. With the accumulation of the zoological samples during a long period, so-called monitoring collections, it becomes possibly to trace the alteration of the marine ecosystems under global climatological, local hydrological and anthropogenic influences.

The creation of electronic databases, firstly on marine invertebrates, started at the Zoological Institute in 1987. Since 1989, PCs helped us in resolving some problems of building and updating data bases and information retrieval systems and allowed us to use them more efficiently.

The lack of a universal international approach to the management of collection data and a number of existing software on the basis of the different computer models, along with some specific problems with data input (*e.g.* Cyrillic symbols), did not allow us to already use compiled foreign software.

The databases on different groups of animals are often interactive and successfully supplement each other. Such combined databases on parasites and their hosts (mammals and fish), insects and food plants etc. were developed at the Zoological Institute. Another example of such a combination is the databases on marine fishes and invertebrates. The material on both groups of animals can be collected as one sample at the same stations and by the same gear. This is the basis for working towards a combined strategy of data input for marine hydrobiology and ichthyology. In 1991, work on ichthyological databases was started (Voronina *et al.*, 1999). Fishes and invertebrates are the main components of any marine biocoenosis and therefore parallel research using joint databases is highly promising for the study of biocoenotic relationships and marine ecosystems.

Designed at the Zoological Institute, the information retrieval system "OCEAN" consists of four main tables: the taxonomic table containing the name and nomenclature of the taxa, the geographical table including the data of field books and catalogue of museum collections (locality of sampling: coordinates of stations, gear etc.), the ecological table (biomass, depth, temperature, salinity, oxygen etc.) and a bibliographic table. The system was improved by a new method of data input with the help of a thesauri system, developed by A.A. Golikov in FoxPro for Windows that minimizes the number of errors.

In spite of fast evolving information systems, standardization and digitizing of biological, in particular zoological, investigations is very slow, partly because of the complications of nomenclature and taxonomical relations. In the course of the organization of faunistic and ecological data and the information retrieval system, two serious problems appeared:

- the input and use of scientific names, especially synonyms
- the formalization of geographical data. The first problem is being solved by using the classifier of scientific names of animals based on the user-friendly and periodically updated ZOOCOD standard, popular among Russian biological institutions dealing with biodiversity research (Table I). The standard was developed in the late 1980s at the Zoological Institute RAS to transform the hierarchical classifications into a relational table (Lobanov and Zaitsev, 1993; Lobanov and Smirnov, 1997; Lobanov *et al.*, 1999).

Coordinates and the developed geographical information system were used to solve the second problem (Dianov and Lobanov, 1995).

The databases comprise information on field stations – localities of collection of marine invertebrates and fish, *i.e.* coordinates, depth, type of bottom, as well as method, gear, date of collecting and collector's name.

The system of the geographical data input takes into account the data standard Darwin Core. In combination with the taxonomic table (information on structure of fauna of certain region) and the collection table (place and method of storing collected specimens), the station data base allows creating different analytical queries to the derivative tables with consideration for hierarchical relations of fish and invertebrate taxa and geographic regions.

Table I. A structure of a ZOOCOD's classificatory system.

GENUS	**LATNAM**	**SYN**	**RANCOD**	**ABBR**	**SYSCOD**
Latin name of genus (not obligatory)	latin name of taxon	code of synonymy	taxonomic code of a rank	unique mnemonic code of taxon	digital systematic code
	Animalia		1	AN	100
	Arthropoda		10	AR	110
	Crustacea		20	CR	120
	Insecta		20	IN	130
	Coleoptera		40	INCO	13010
	Diptera		40	INDI	13013
	Chordata		10	CH	140
	Mammalia		20	MA	150
	Primates		40	MAPR	15010
	Pongidae		50	MAPRPO	15010100
	Hylobatidae	=	50	MAPRHY	15010100
	Gorilla		70	MAPRPOGOR	150101001000
Gorilla	gorilla		90	MAPRPOGORGOR	1501010010001000
	Pan		70	MAPRPOPAN	150101001010
	Hominidae		50	MAPRHO	15010105
	Homo		70	MAPRHOHOM	150101051000
Homo	sapiens		90	MAPRHOHOMSAP	1501010510001000
Homo	recens		94	MAPRHOHOMSAPRE	150101051000100010

The creation of the electronic databases and the design of the information retrieval systems OCEAN and ZOOINT (in standard ZOOCOD) carried out at the Zoological Institute, have allowed us to receive support for the project entitled "ECOANT" - "Creation of an information retrieval system on ECOlogy of benthos of the ANTarctica". The information retrieval system "ECOANT" can promote the resolution of the following problems:

- to refine the faunal structure of biota and its taxonomic features for the different areas
- to obtain ecological information
- to reveal changes in the structure of fauna in the investigated regions under influence of climatological and anthropogenic factors, which is one of the aims of the global ecological monitoring. The realization of the project is based mainly on the Russian biological data of Antarctic Regions and, first of all, on the unique benthic collections of the Antarctic and Sub Antarctic Seas. The preliminary information about this project is available on the web (http://www.zin.ru/projects/ecoant/index.html).

Using the Active Server Pages technology the database on the Antarctic seabirds, brittle stars and chitons are presented on the server of the Zoological Institute (http://www.zin.ru/projects/ecoant/eco1form.asp). The list of fishes and fishlike vertebrates of the Antarctic Region is prepared to come online.

The results of developing and updating the information retrieval system "OCEAN" on fauna of the Arctic, Antarctic, Far East Seas and inland seas of Russia will be unique since the collections of marine fishes and invertebrates, accumulated and kept by generations of scientists during almost two ages, serves as a unique source of information.

By now the ichthyological part of the taxonomical table contains over 5,800 records: all high-level taxa including families, taking into consideration modern fish taxonomy; species of fish and fishlike vertebrates of the Antarctic Region, flatfishes (order Pleuronectiformes - 13 families) and scorpaeniform fishes (suborders Scorpaenoidei - 10, Cottoidei - 3 and Platycephaloidei - 2 families) of the world. The collection table includes data on fish specimens of the orders Pleuronectiformes and Scorpaeniformes as well as some fish species of the Antarctic Region kept at the Zoological Institute RAS. The database on marine invertebrates contains over 15,000 station records and information on chiton, bivalve and brittle star collections. Some characteristics of the collection database on marine fish and invertebrate collections are presented in Table II.

The information of the collection database is, however, still incomplete, because it covers information only for some taxa. Nevertheless up to now it is already possible to use the collection databases in analysis of secondary information and it will show some characteristics of the fish and invertebrates collection.

Table II. Some characteristics of databases on fish and invertebrates collections.

Group of animals	**Number of stations**	**Number of taxa**	**Number of inventory units**	**Specimens**
Fish	5,845	272 genera, 710 species	9,083 (including 157 types)	26,524 (including 23 stuffed fishes)
Invertebrates (Arctic)	14,897	62 genera, 110 species	11,913	82,851 wet, 19909 dry
Invertebrates (Antarctic)	2,520	64 genera, 136 species	2,887	2,619 wet, 1,608 dry

A great part of the material of the Zoological Institute was collected during the well-known Russian expeditions as well as foreign ones (Table III). Only few expeditions published their route and station data in the special issues and for many years these works were not available, even for specialists (Lindberg, 1954). Many other expeditions have only handwritten diaries. Sometimes the label data, and therefore catalogue data, are very fragmentary. The creation of the joint international expedition database would be historically very interesting and also very useful for further input of collection data in helping to unify data and in avoiding errors during input. The example of such a useful information source is the Challenger Expedition 1873-1876 database on the site “Biogeoinformatics of the Hexacorals: http://www.kgs.ku.edu/Hexacoral/” (Fautin and Buddemeier, 2003).

Table III. The most extensive collected material of Russian and foreign expeditions kept at the ZIN.

Name and abbreviation of expedition	Number of "ichthyological" stations	Number of "invertebrate" stations	Date
Polar Exp. of K. Baer	12		1840
Murman Scientific-Fishery Exp. (ENPIM)	482	2896	1880-1915
Spitsbergen Exp.	34	2	1899-1901
Russian Polar Exp. (RPE)	44	107	1900-1902
Novaya Zemlya Exp.	35	13	1901-1935
Baltic Exp.	23		1907-1908
Far East Exp. (FEE)	388	81	1908-1915
Hydrographic Exp. To East Ocean (HEEO)	128	279	1908-1927
Exp. ZIN to Japan Sea	63	32	1934
VNIRO Kara Sea Exp.	46	19	1945-1946
Kuril-Sakhalin Exp. (KSE)	314	1156	1947-1949
Soviet-Chinese Exp. (SCE)	37	16	1956-1959
Southern Sakhalin Exp.	40	12	1946
TINRO Exp.	68	173	1928-1978
Tropic Exp.		147	1974-1975
Arctic Exp. "Polarstern" (Germany)	18	222	1985-1998
MERA-95		194	1995
Severnyi Polyus (North Pole)	19	404	1946-1948
Shantar Exp.		300	1978
Antarctic expeditions			
Antarctic Exp. "Polarstern" (Germany)		111	1972-1978
AzCherNIRO Exp.		336	1969-1976
Soviet Antarctic Exp. (SAE)		843	1956-1989

An example of one of the most intensive expansions of collections, in relation with the long and extensive expeditions such as ENPIM (1880-1915) is given in Fig. 1. It is expected that the periods of severe social circumstances (1917-1920 and 1941-1945) are characterized by very few samplings of zoological material.

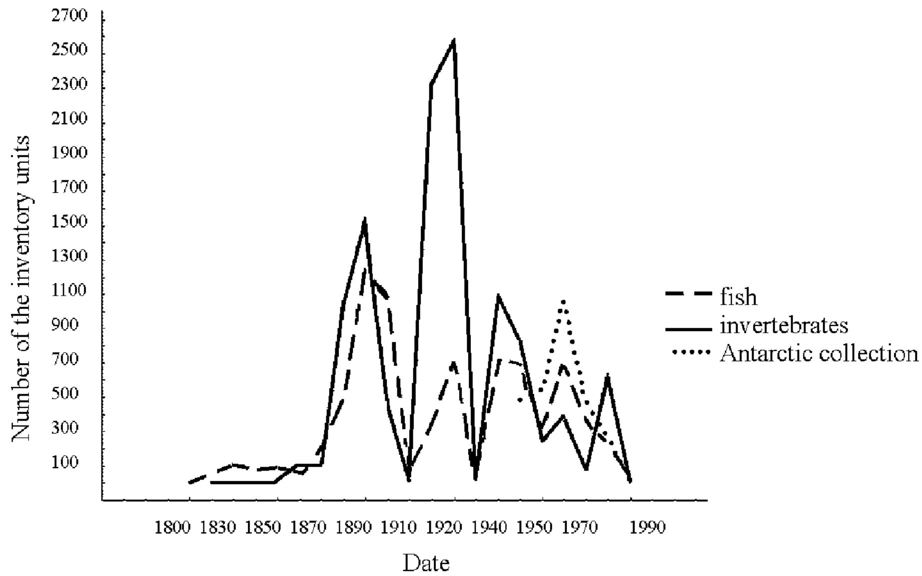


Fig. 1. Chronological chart of the fish and invertebrates collecting (according to data bases today).

In total about 50 expeditions and 850 collectors sampled zoological material since 1828 (Table III, IV) aboard of 180 vessels, among them some well-known scientific vessels such as - "Andrei Pervozvannyi"; "Vityaz"; "Akademik Knipovich" (VNIRO); "Ob" (AARI); "Skif" and "Aelita" (AzCherNIRO); "Zund" and "Evrika" (AtlantNIRO) as well as Marine Fishery Fleet vessels, and occasionally even military vessels. The part of the samples collected by the academician vessels is 5-10%. The biggest part of the benthic collection (65-70%) was sampled during expeditions on board Marine Fishery Fleet vessels, despite of defects of the sampling and labels, difficulties of preservation and storage etc. Each cruise, even those with little material, contributed to faunistic investigations of this unique world of marine life.

Table IV. The names of the collectors of the largest number of samples recorded in our current databases.

Name of collector	Number of ichthyological stations	Number of invertebrate station
Andriashev A.P.	96	
Arngold E.E.	26	114
Averintsev V.G.	12	270
Barsukov V.V.	127	
Brazhnikov V.K.	71	
Bryazgin V.F.	5	395
Bunge A.A.	47	
Byalynitskii-Birulya A.A.	10	135
Bykhovskii B.E.	66	
Derbek F.A.	92	
Fedorov V.V.	106	
Foroshchuk V.P.	97	
Golikov A.N.	20	303
Gorbunov G.P.	55	447
Gruzov E.N.	7	89
Gurjanova E.F.	43	
Herzenstein S.M.	20	296
Knipovich N.M.	27	481
Kobyakova Z.I.	4	119
Koltun V.M.	12	513
Kondakov A.N.	5	83
Legeza M.I.	90	
Lindberg G.U.	93	
Neyelov A.V.	45	
Petryashov V.V.		256
Rutenberg E.P.	78	
Shmidt P.Ju.	200	
Sideleva V.G.	97	
Sirenko B.I.	5	557
Smirnov A.V.	20	197
Soldatov V.K.	364	
Starokadomskii L.M.	36	123
Ushakov P.V		333
Vagin V.L.	65	368
Vinogradov L.G.		310
Voznessenskii I.G.	31	

Some material (*e.g.* 178 inventory numbers of the ichthyological collection and more than hundreds invertebrates) has been received in exchange with foreign museums.

The text catalogues of the collections of the Zoological Institute were published only for type specimens. The catalogue of all flatfish collections (Pleuronectiformes) has been compiled on the basis of the information retrieval system OCEAN (Voronina and Volkova, 2003) and published recently. It is planned to launch the collection databases directly on the internet.

Table V. Comparison of the information retrieval systems OCEAN and ARTEDIAN.

Advantages of OCEAN	**Advantages of ARTEDIAN**
Station unique codes are generated automatically.	By default more fields describing the object can be filled in
Leading spaces are being deleted automatically.	The collectors are treated in separate fields.
The important fields (vessel and others) are filled in Russian, in addition to English.	The IRS is conform the Darwin Core in coordinate notation (degrees, minutes and seconds separately)
Some more additional fields (EXPEDITION, GEAR, STATION NUMBER etc.) provide an opportunity to verify location with a route table for marine expeditions.	Additional fields are designed for freshwater stations, conform the Darwin Core (lakes, rivers, states, provinces, county etc.)

International projects to create data base and information retrieval systems for sharing biodiversity information on a global scale are in development. Some examples of the comparison of the data input systems of OCEAN and Artedian, the system used in creating the collection databases FishBase, is presented in Table V. The main point is that in foreign projects the system of data input is exclusively based on Latin symbols, but not Cyrillic. This restricted approach leads to a considerable reduction of information originating from the actual labels that are often hand-written in local national languages, *e.g.* Russian. In addition, it is sometimes difficult to translate these data equivalently and completely. Therefore we feel it is important to consider special fields in the tables of the information retrieval system to allow input of data as well as to perform the queries in Cyrillic symbols together with the Latin ones. The information retrieval system OCEAN provides this possibility. It is also worth noting the necessity to examine the experiences of different teams and try to create not ideal but optimal systems for input and treatment of marine data.

Acknowledgements

Supported came from the Project N11 "Exploration and research of the Antarctic Region" of the Federal Program "World Ocean", grants № 05-07-90354, 04-04-49300, NSh 1668.2003.4 and program "Information system on a biodiversity of Russia".

References

Atlas of Antarctic. 1969, v. 2. Gidrometeoizdat, L. 598p.

Dianov M.B. and A.L. Lobanov. 1995. Computerized geographical system ZOOMAP for mapping of plants and animals areas. In: Abstr. II Soveshchanie "Kompyuternye bazy dannykh v botanicheskikh issledovaniyakh". St. Petersburg, April 17-19 1995:16-17

Fautin, D.G. and R.W. Buddemeier. 2003. Biogeoinformatics of the Hexacorals: http://www.kgs.ku.edu/Hexacoral/

Lindberg G.U. 1954. Obzor rabot Kurilo-Sakhalinskoi morskoi kompleksnoi ekspeditsii Zoologicheskogo Instituta I Tikhookeanskogo instituta rybnogo khozyaistva (Review of the works of the KSE). Trudy Kurilo-Sakhalinskoi Ekspeditsii ZIN-TINRO, 1947-1949. Vol. 1. M.-L.:7-100.

Lobanov A.L., I.S. Smirnov and M.B. Dianov. 1999. ZOOCOD – conception of representation of zoological hierarchical classifications in relational databases. In: Abstracts of the Int. Symposium "Information retrieval system in biodiversity research", Proceedings of the Zoological Institute, V. 278. St. Petersburg:65-66.

Lobanov A.L. and I.S. Smirnov. 1997. Principles of arrangement and using of classifiers of animals in the standard ZOOCOD. In: Data bases and computer graphics in zoological investigations. Proceedings of the Zoological Institute, V. 269. St. Petersburg:66-75.

Lobanov A.L. and M.V. Zaitsev. 1993. Creation of computer data bases on the systematics of mammals on the basis of classificatory of animals names "ZOOCOD". In: Questions of systematics, faunistics and paleontology of small mammals, Proceedings of the Zoological Institute. V. 243. S.-Petersburg:180-198.

Smirnov I.S., A.L. Lobanov and M.B. Dianov. 1999. Zoological digital (virtual) museums. In: Scientific service in network Internet. Abstract of All-Russian scientific conference in Novorossiysk on September 20-25, Publ. of Moscow University:185-187.

Smirnov I.S. and A.V. Neyelov. 1996. Studying of Antarctic bottom fauna in cruises of fishery vessels of USSR and Russia. In: Abstr. of Int. conference: The history of native oceanology. Kaliningrad, 28 October – 1 November 1996:106-107.

Voronina E.P. and G.A. Volkova. 2003. Catalogue of specimens in the collection of the Zoological Institute, Russian Academy of Sciences. Osteichthyes, Pleuronectiformes. Explorations of the fauna of the seas. Vol. 55(63). St. Petersburg, Zoological Institute RAS. 198p.

Voronina E.P., I.S. Smirnov and A.A. Golikov. 1999. The computer approaches to the ichthyological studies in Zoological Institute RAS. In: Abstracts of the Int. Symposium "Information retrieval system in biodiversity research", Proceedings of the Zoological Institute, V. 278. St. Petersburg: 116-117.

Identifying erroneous data using outlier detection techniques

Wei Zhuang[1], Yunqing Zhang[2] and J. Fred Grassle[2]

[1]Department of Computer Science, Rutgers, the State University of New Jersey, Piscataway, NJ 08854-8019, USA E-mail: weiz@paul.rutgers.edu

[2]Institute of Marine and Coastal Sciences, Rutgers, the State University of New Jersey, New Brunswick, NJ 08901, USA

Abstract

Common data quality problems observed in OBIS data integration processes are described. DBSCAN, a density-based clustering algorithm for large spatial databases is employed to identify geographical outliers in federated data from a public Web service on the OBIS Portal. The algorithm is shown to be effective and efficient for this purpose. The relationship between outliers and erroneous data points are discussed and the future plan to develop an operational data quality checking tool based on this algorithm is discussed.

Keywords: QA/QC; outliers; clustering; data quality solving.

Introduction

Federated scientific databases such as the Ocean Biogeographic Information System (OBIS, http://www.iobis.org) and the Global Biodiversity Information Facility (GBIF, http://www.gbif.org), have solved the data heterogeneity problem by employing open communication protocols and data exchange standards (Grassle and Stocks, 1999; Zhang and Grassle, 2003). For the first time in science history, tens of millions of records about our shared biodiversity heritage have been made publicly accessible on the World Wide Web. Although scientists and data managers have carefully performed quality checking over individual datasets and collections, data corruptions can still occur during data compilation, re-entry, conversion, and the transfer process. For example, default null database objects can turn into string “null”s; latitudes and longitudes are reversed; non-ascii characters are mistakenly encoded, etc. These data quality concerns have been familiar to data warehousing communities and a great deal of research and development has been carried out in this area. Our problem, which is data quality checking over federated ocean biodiversity information, is unique in at least two aspects:

- DQ solving has to be efficient for large spatial datasets
- The domain knowledge is highly specialized and may not be translated into simple database constraints in many cases (this is indeed a common problem for scientific data management).

Outliers are commonly defined as rare or atypical data objects that do not behave like the rest of the data. Often, erroneous data points appear as outliers in a database. Scientists and data managers have used visualization tools to identify outliers in datasets. When the dimension of the feature space is more than two, visual identification becomes challenging. Moreover, when the database multiplies in content, manual identification by naked eyes becomes infeasible. Henceforth what is needed here is an automatic outlier detection tool that can efficiently handle large, high dimensional databases. It should be made clear that automatic tools are not to replace domain scientists' opinion in data quality checking, rather, they are "pre-processors" to provide assistance to domain scientists. The question of how to integrate domain knowledge in automatic outlier identification tools will be discussed elsewhere. In this paper we will concentrate on the algorithm testing aspect of the development.

Section II describes the method. In section III we report and analyze the results. We discuss the results and direction for future work in Section IV.

Method

There is a considerable body of research on outliers by statisticians (Barnett and Lewis, 1994; Hawkins, 1980). Fitting databases with parametric models requires prior knowledge of data distribution and using parametric models in the data processing stage may lead to circular arguments and produce spurious patterns when doing data analysis. Non-parametric clustering algorithms are attractive for grouping objects in a database into subclasses and, intuitively, small clusters, or classes with few members, are where outliers are. Computer scientists have been conducting extensive research to develop efficient clustering algorithms for large databases (Berkhin, 2002; Guha *et al.*, 1998; Zhang *et al.*, 1996; Ng and Han, 1994,).

The well-known K-means algorithms partition a dataset into a set of k clusters in two steps: firstly it determines the k cluster centers by minimizing an object function; secondly it assigns a cluster membership based on the distance of the data object to the cluster centers. In these partition-based algorithms, the number of clusters, k, is an input parameter provided by the user while in many cases the user has no idea of the number of clusters. These algorithms are sensitive to noise. We then investigate a different, density-based family of clustering algorithms. In these algorithms, parts of the feature space with dense data points form clusters while outliers have a much lower density and are further away from the clusters. DBSCAN (Ester *et al.*, 1996), DENCLUE (Hinneberg and Kleim, 1998) and WaveCluster (Sheikholeslami *et al.*, 1999) are well known algorithms in this family. It has been demonstrated that DBSCAN requires minimal domain knowledge, can discover clusters with arbitrary shapes and is efficient on large databases. Most importantly, it can separate "noise" (*i.e.* outliers) while performing clustering. Details of the algorithm can be found in Ester *et al.* (1996).

We obtained the software from the first author as a C++ package and adapted it for OBIS data. The clustering was run on a Sun Solaris 9 machine. The experimental data are provided by the OBIS portal as a Web service. We tested the algorithm by identifying geographical outliers and great circle distance is used to define the distance function between two data points.

Results

Here we report the results for three species: *Euthynnus alletteratus*, *Albula vulpes* and *Balaenoptera borealis*. The time complexity of DBSCAN is O (nlogn) where n is the number of data points. In table I we list the run time for these three experiments and the number is consistent with the O (nlogn) estimation.

Table I: Runtime for clustering and identifying outliers using DBSCAN.

Dataset	Number of records	Runtime (in milliseconds)
Euthynnus alletteratus	338	1780
Albula vulpes	840	5693
Balaenoptera borealis	7125	424910

In Fig. 1-3 we show clustering results for the three species where outliers are represented by round dots and non-outliers triangles. Examining the three figures together with the underlying datasets, we see that this algorithm correctly identifies all the single records far away from data clusters. Some non-outliers may appear to be outliers to the naked eye. For example, the triangle at (44.15°N, 6.03°E) in Fig. 2 is far away from the other clusters but in fact it represents 12 individual data records and thus is not an outlier in its common definition. One could visit the OBIS Portal to look up the interactive maps and download the datasets for further confirmation of our results.

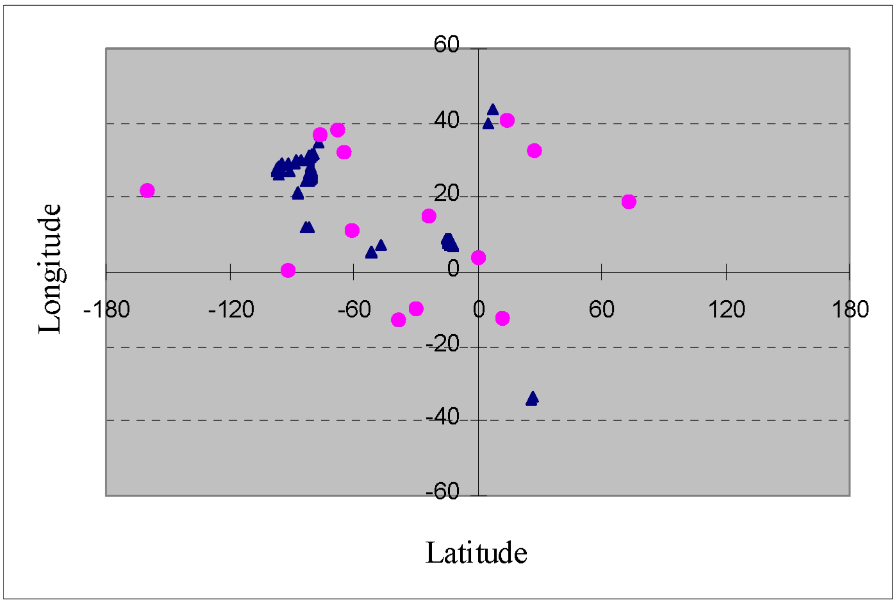


Fig. 1. Result set for Euthynnus alletteratus.

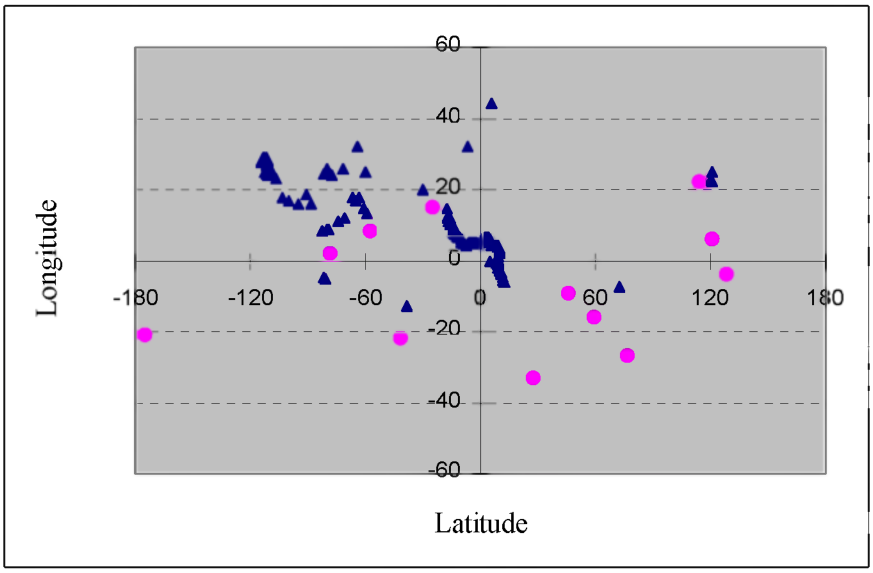


Fig. 2. Result set for Albula vulpes.

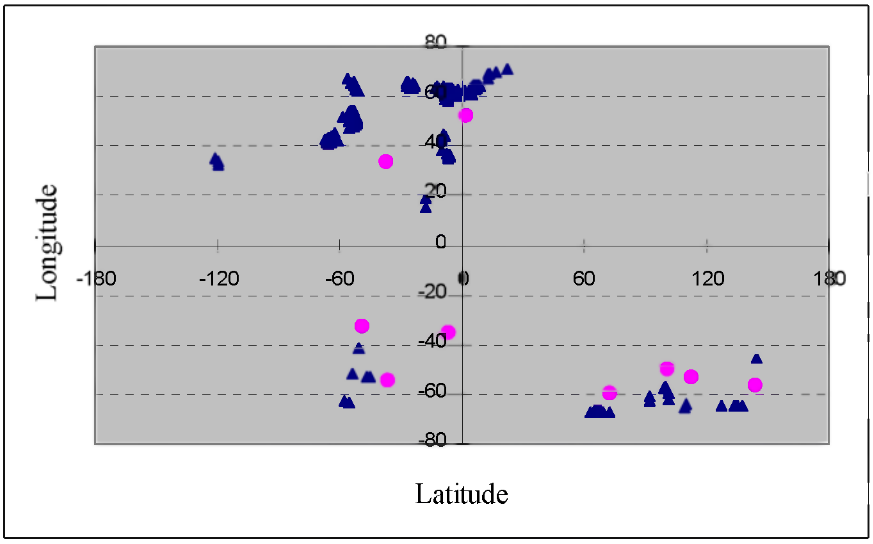


Fig. 3. Result set for Balaenoptera borealis.

Discussion and Future Direction

In this work we have demonstrated the usability of the density-based clustering algorithm —DBSCAN— in identifying geographical outliers. Because sampling is not complete yet, the outliers are not necessarily erroneous data points. Sometimes they are rare sightings or a single specimen in museum collections. Under these circumstances an expert has to examine the outliers and identify the actual erroneous data. On the other hand, the other features in the data space (temperature, salinity, etc…) may have been better sampled and outliers identified in those feature spaces may be more an indicator of data errors. In fact we have performed preliminary studies on temperature space and the results are promising. In the next step, we will develop an incremental learner where outlier detection results obtained from different feature spaces are combined. Domain scientists will play an active and critical role in this learner because:

- they will be prompted with candidates produced by the outlier detection program and select the erroneous data from the candidates
- their decision will be fed back to the learner where the relative weights assigned to individual learners will be readjusted.

Conclusion

The clustering algorithm —DBSCAN— has been successfully applied to identifying geographical outliers in OBIS point data on a species-by-species basis. The algorithm is efficient enough to scan large spatial databases such as OBIS. With more samples coming into OBIS, the outlier detection technique can be used to identify erroneous data points and be part of an operational data quality checking tool where domain knowledge and automatic learners are integrated in a dynamic way.

References

Barnett V. and T. Lewis. 1994. Outliers in Statistical Data. John Wiley & Sons, Chichester, New York. 608p.

Berkhin P. 2002. Survey of Clustering Data Mining Techniques. Accrue Software. http://citeseer.ist.psu.edu/berkhin02survey.html.

Ester M., H.-P Kriegel., J. Sander and X. Xu. 1996. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, 226-231.

Grassle J.F. and K.I. Stocks. 1999. A Global Ocean Biogeographical Information System (OBIS) for the Census of Marine Life. Oceanography 12(3):12-14.

Guha S., R. Rastogi and K. Shim. 1998. Cure: An Efficient Clustering Algorithm for Large Databases. Proc. of ACM SIGMOD Int'l Conf. on Management of Data, 73-84. ACM Press

Hawkins D. 1980. Identification of outliers. Chapman and Hall, London. 188p.

Hinneburg A. and D.A. Kleim. 1998. An Efficient Approach to Clustering in Large Multimedia Databases with Noise, KDD'98, New York, Aug. 1998.

Ng R.T., J. Han. 1994. Efficient and Effective Clustering Methods for Spatial Data Mining 20th International Conference on Very Large Data Bases, September 12-15, 1994, Santiago, Chile, Proceedings. 144-155.

Sheikholeslami G., S. Chatterjee, A. Zhang. 1999. WaveCluster: A Wavelet Based Clustering Approach for Spatial Data in Very Large Databases. 289-304.

Zhang T., R. Ramakrishnan, M. Livny. 1996. BIRCH: An Efficient Data Clustering Method for Very Large Databases. ACM SIGMOD International conference on Management of Data, Montreal, Canada. 103-114.

Zhang Y. and J.F. Grassle, 2003. A Portal for the Ocean Biogeographic Information System, Oceanologica Acta. 25(5):199-206.