Vol. 316: 201–202, 2006 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published July 3

THEME SECTION

Ocean biodiversity informatics (OBI)

Idea and coordination:
Mark J. Costello, Edward Vanden Berghe, Howard I. Browman

CONTENTS

Introduction

Mark J. Costello[1],*, Edward Vanden Berghe[2], Howard I. Browman[3]

[1]Leigh Marine Laboratory, University of Auckland, PO Box 349, Warkworth, New Zealand
[2]Flemish Marine Data and Information Centre, Flanders Marine Institute, Wandelaarkaai 7, 8400 Oostende, Belgium
[3]Institute of Marine Research, Austevoll, 5392 Storebø, Norway

The computerised information age is providing new opportunities and challenges for marine science, ranging from electronic publication (Kinne 1999) to interoperable databases that provide access to primary data over the Internet. This Theme Section (TS) describes some of the activities in the field of ocean biodiversity informatics (OBI), whereby information technologies are used to support the management of data and information on ocean biodiversity. The first contribution is a review of the subject (Costello & Vanden Berghe), the next 5 contributions provide examples of internationally important, cutting-edge marine information sys-

*Email: m.costello@auckland.ac.nz

tems that are available online (Fabri et al., Arvantidis et al., Lleonart et al., Halpin et al., Stevens et al.), and the last 3 contributions are analyses that demonstrate the benefits of easy access to such large databases (Costello et al., Guinotte et al., Kaschner et al.).

This TS was stimulated by a series of meetings developed by the growing international cross-disciplinary community of marine biologists and data managers. A workshop sponsored by the Intergovernmental Oceanographic Commission (IOC) of UNESCO brought physical oceanographers, biologists and data managers together in 1996; it was followed by a symposium on ocean data management in 2002 (Vanden Berghe et al. 2004; available at: www.vliz.be/En/activ/events/cod/cod.htm). An international conference on Ocean Biodiversity Informatics (29 November to 1 December 2004) attracted >170 delegates from 37 countries, and 70 presentations (available at: www.vliz.be/obi). Participants came from the Global Biodiversity Information Facility (available at: www.gbif.org), government agencies, universities, NGOs, museums, and commercial companies, demonstrating the breadth of organisations and expertise involved in OBI. Most of the contributions to this TS are based upon presentations at this conference.

One trend in OBI is that central databases are being replaced by online data systems that make both primary and secondary data freely available. For example, the leading provider of primary data on marine species distributions, the Ocean Biogeographic Information System, has grown rapidly in the past 5 yr, demonstrating how marine species data can be shared from local to global scales, and mapped with ocean environment information. This revolution in data availability provides new opportunities for marine science. There are no longer any technical or data availability obstacles to the inclusion of marine biological data in information networks that focus on physical ocean and climate data (e.g. national and world ocean data centres, Global Ocean Observing System, Global Earth Observation System of Systems). In addition to the publication of synthesized data in standard journals, scientists can now publish their primary data, so that other researchers may build on these to provide added value. This data exchanges benefits from common vocabularies and protocols that will better define scientific concepts and facilitate better understanding of marine ecosystems at all spatial scales.

Acknowledgements. The 2004 Ocean Biodiversity Informatics conference was sponsored by the Intergovernmental Oceanographic Commission (IOC) of UNESCO, and IOC's International Ocean Data and Information Exchange; the International Council for Exploration of the Sea; the Census of Marine Life's Ocean Biogeographic Information System; the International Association of Biological Oceanography; the Taxonomic Data Working Group (http://www.nhm.ac.uk/hosted_sites/tdwg/); Flanders Marine Institute; the European Marine Biodiversity and Ecosystem Function research network (MarBEF); the European Commission and the German Government. H.I.B.'s research, and his editorial activity for MEPS, are supported by the Institute of Marine Research, Norway, The Research Council of Norway, and the Inter-Research Science Centre.

LITERATURE CITED

Kinne O (1999) Electronic publishing in science: changes and risks. Mar Ecol Prog Ser 180:1–5

Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) (2004) The colour of ocean data. Proc Symp, Brussels 25–27 November 2002. IOC Workshop Report 188, UNESCO, Paris, Vlaams Institut voor de Zee Spec Publ no. 16

Vol. 316: 203–214, 2006

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

'Ocean biodiversity informatics': a new era in marine biology research and management

Mark J. Costello[1,*], Edward Vanden Berghe[2]

[1]Leigh Marine Laboratory, University of Auckland, PO Box 349, Warkworth, New Zealand
[2]Flemish Marine Data and Information Centre, Flanders Marine Institute, Wandelaarkaai 7, 8400 Oostende, Belgium

ABSTRACT: Ocean biodiversity informatics (OBI) is the use of computer technologies to manage marine biodiversity information, including data capture, storage, search, retrieval, visualisation, mapping, modelling, analysis and publication. The latest information systems are open-access, making data and/or information publicly available over the Internet. This ranges from primary data on species occurrences, such as in the Ocean Biogeographic Information System (OBIS), to species information pages and identification guides. Using standard data schema and exchange protocols, online systems can become interoperable and, thus, integrate data from different sources. However, insufficient metadata standards, i.e. the terminology to describe data, are available for biology and ecology. Quality assurance needs at least the same rigour as for printed publications, including expert oversight (e.g. Editorial Board), quality-control procedures and peer review. An index of data use is proposed to parallel citation indices for printed journals. Other challenges include data archiving and Internet access in developing countries. Although taxon names are the central, and most unique, element of biodiversity informatics, only about one-third of the names of described marine species are currently available online in authoritative master lists. The scientific community can form alliances that build and maintain biodiversity informatics infrastructures and that address data ownership and commercialisation potential. OBI enables greater access to more data and information faster than ever before, and complements the traditional disciplines of taxonomy, ecology and biogeography. It is urgently needed to help address the global crises in biodiversity loss (including fisheries), climate change and altered marine ecosystems. For OBI to succeed, governments, science-based organisations, scientists and publishers need to insist on online data publication in standard formats that enable interoperability. This change in marine biology culture is already underway.

KEY WORDS: Data schema · Data exchange protocols · Interoperability · Archiving · Quality assurance · Peer review · Nomenclature · Taxonomy · Biogeography

INTRODUCTION

For several hundred years marine biology has been based on natural history, and during the 20th century began to address ecology and evolution. In recent decades, genetic and molecular sciences have brought new insights to marine biology. In parallel, physical oceanography has become a global science that uses satellites and other remote-sensing technology to complement traditional sampling. Plans for real-time sharing of data are underway as part of the Global Ocean Observing System (GOOS). This growth in physical data led to the Intergovernmental Oceanographic Commission's (IOC) International Oceanographic Data and Information Exchange (IODE) programme, establishing a network of national ocean data centres (NODC) around the world. While remote and automated *in situ* methods are successful for the frequent

[*]Email: m.costello@auckland.ac.nz

gathering of physical and chemical data over large areas, collecting biological data is more difficult, due to the small body size of most organisms, diversity of species and contrasting habitats where they occur (Fautin & Fippinger 2005). These challenges and related costs involved in collecting biological data make its publication all the more important. However, with the exception of genetic data, marine biology data has remained scattered and often unpublished (Grassle & Stocks 1999, Grassle 2000, Myers 2000, Zeller et al. 2005). This may reflect the lack of opportunities for publication of raw data until recently. The Internet has reduced costs of data publication, and marine biology has entered the information age along with other sciences (Kinne 1999, International Council for Science 2004). In the present paper, we define the scope, challenges and future prospects for the new field of ocean biodiversity informatics (OBI).

Need for data access

Never before has the need for rapid access to data at regional and global scales been so important. Recent analyses of ocean-scale data have shown major shifts in plankton distribution due to climate (e.g. Stevens et al. 2006, in this Theme Section), global over-fishing (Pauly et al. 2003), manifold reductions in abundance of large fish (e.g. Myers & Worm 2003), profound changes in ecosystem structure because of indirect effects of fisheries that may be irreversible (Jackson et al. 2001, Frank et al. 2005), and as yet unexplained 62 million yr cycles of marine genera richness in the 542 million yr fossil record (Rhode & Muller 2005). Without informatics-aided analyses and large-scale databases to support them, the global nature of these phenomena would not have been recognised.

Species are being introduced by human activities around the world, with ensuing socio-economic impacts on local fisheries, aquaculture and human health. Often these species may not be recognised as introductions, because so far only a fraction of marine species have been described. The ability to identify species from anywhere in the world is particularly important for the detection of introductions that may prove economically harmful. Global fisheries statistics reporting was compromised by poor species identification, prompting the FAO to produce species identification guides and fact-sheets (Lleonart et al. 2006, in this Theme Section). Online species identification guides are immediately accessible to people who have Internet access (e.g. www.crustacea.net). In addition, electronic keys helpfully allow users to select whichever characteristics of the animal or plant they can recognise with confidence. In contrast, traditional keys force the user to choose 1 or 2 characters at each step, such that 1 error or oversight can lead to lost time and misidentification. Information systems are built to manage the ever-increasing volume of data available on invaders (e.g. www.issg.org/database, www.gisinetwork.org). Software tools such as the Kansas Geological Survey Mapper (e.g. Guinotte et al. 2006, in this Theme Section) and Desktop GARP (e.g. Wiley et al. 2003), can be used to predict potential environmental suitability for candidate invasive species. Other modelling approaches may be less automated (e.g. Kaschner et al. 2006, in this Theme Section). It seems likely that ocean biodiversity informatics will provide a suite of modelling options appropriate for different types of data and purposes.

Local patterns of biodiversity have their origins in, and may still be maintained by, ecosystem processes that occur at regional and global scales. Thus, selecting areas for fishery stock management and conservation requires knowledge of biodiversity patterns at all spatial scales. At present, conservation too often focuses on national-scale patterns, because of regulatory obligations and the limited availability of data at larger geographic scales. Conservation should, however, operate at ecologically and evolutionarily relevant scales and, thus, requires access to data at a range of spatial scales.

Data, information and knowledge

Data and associated metadata (background information about the data) are the foundation of science; the what, where, when, who and how. The interpretation of these facts leads to information and theories that create knowledge. At present, marine biology delivers many papers that provide statistics, graphs and models derived from often unpublished data. While the importance of most of these syntheses, models and theories will eventually fade, the value of the data increases with time, as it becomes harder to replace. The digitisation of historical data from paper files can cost ≤0.5% of that of the original field surveys (Zeller et al. 2005), and can reveal new insights into human interactions with natural resources (e.g. Lotze & Milewski 2004).

Scope of ocean biodiversity informatics

Biodiversity informatics is the computer technology that enables the management and analysis of biodiversity data and information (Bisby 2000); it has many benefits and positive outcomes (Table 1). The Convention on Biological Diversity definition of biodiver-

Table 1. Some of the benefits of biodiversity informatics

Data publication
- Low cost publication of text, maps, images, movies, sounds
- Easier access to data and metadata
- Availability of data and metadata widened
- Rapid publication
- Linking to many data and information resources on the world wide web

Consequences
- Permits data mining and exploration
- Combination and sharing of data from multiple sources
- Data are re-usable for perhaps unforeseen benefits
- Repatriate data and knowledge to developing countries

- Interactive and/or user-defined readability
- Data management tools widely available at little to no cost
- Automated calculation of statistics (e.g. how many species, hotspots, gap analysis)
- Demonstrates good quality data management
- Gaps in data and information are more apparent

- Collaboration between different research groups is promoted and facilitated
- Awareness of the localities and collections where species occur facilitates researchers to visit them
- Non-biodiversity researchers may analyze the data from new perspectives

- Policy makers and the public can become more engaged by having transparent access to the data from which conclusions have been drawn
- Increase public confidence in a more transparent and accessible science
- Improved training and education because teachers can obtain real data sets for student exercises

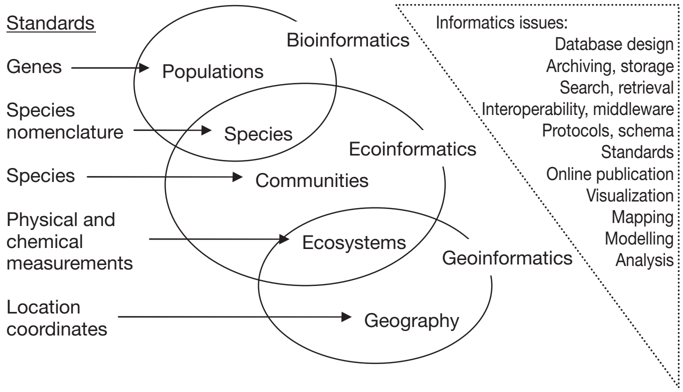


Fig. 1. Diagram showing how 3 sub-components of biodiversity informatics relate to aspects of biodiversity from genes to ecosystems and the environment. Each component has 1 unique aspect and also areas that overlap with others. The essential and standard fields for each aspect and informatics issues common to all components are indicated

sity covers the variety of life within species (e.g. populations), between species (e.g. communities) and of ecosystems (i.e. ecological and environmental interactions) (Costello 2001). Related fields include bioinformatics, phyloinformatics, species informatics, ecoinformatics and geoinformatics (Fig. 1). The term 'bioinformatics' is generally restricted to molecular and genetic data that do not involve species names as a core element. 'Phyloinformatics' concerns the phylogenetic relationships between taxa (e.g. Tree of Life initiative; www.tolweb.org). Species, eco- and geoinformatics concern species level, ecological and ecosystem, and geographic aspects, respectively. They focus on concepts described as words, notably species, habitats and places, respectively, rather than numerical or biochemical data. It is to these text-based concepts that biodiversity informatics provides the most novel contributions and solutions.

OBI is an interdisciplinary activity based on data associated with marine species and their environment. It includes traditional database design and function, as well as data exchange standards, schema and protocols, and exploration, visualisation, analysis and publication software (Fig. 1). While primary goals are free- and open-access to data over the Internet, some project-specific or sensitive data (e.g. location of threatened species) may be withheld. The use of open-source software is preferred (e.g. Linux, MySQL, MapServer), because this can be modified for special purposes and freely shared, but standard proprietary software is also used (e.g. Oracle, Microsoft Access, ARCIMS). Chapman (2005b) lists examples of 18 software resources for biodiversity data management, modelling, georeferencing and mapping, quality control and data analysis; and most are free.

STANDARDS

With the advent of online data exchange, standard data exchange protocols, middleware (or wrappers) that cross-map one database to another, and common vocabularies of terminology are now more in demand than before, when databases were isolated and centralised. Standard categories and definitions are also required for the metadata that describes datasets ('discovery metadata') and data records. Whereas links between web pages are by hypertext mark-up language (HTML), the extensible mark-up language (XML) and the development of ontologies provide a more formal structure for data exchange protocols (Millard 2004, Reed & Pissierssens 2004). Standards have been developed by the International Standards Organisation (ISO; www.iso.org) for many aspects of environmental data management. Unfortunately, the reports describing the standards are not available free of charge, which limits their widespread use.

Data exchange

A standard list of data fields (48 data elements) for exchanging data on species distribution records called 'Darwin Core' has been established. This has been expanded in a backward-compatible manner by the Ocean Biogeographic Information System (OBIS) and the Mammal Networked Information System (MaNIS) for marine and mammal specialisations, respectively. In biology, the most widely used data access protocol is DiGIR (Distributed Generic Information Retrieval). The Access to Biological Collections Data (ABCD) schema is more complex and comprehensive (about 700 data elements) than Darwin Core, and has become a TDWG (Taxonomic Data Working Group; www.tdwg.org) standard. Since ABCD has a hierarchical structure and incorporates repeatable elements (e.g. for multiple determinations, images), the BioCASE protocol had to be developed to transport ABCD data. A protocol is under development which builds on and combines DiGIR and BioCASE. This is called TAPIR (the TDWG Access Protocol for Information Retrieval; www3.bgbm.org/protocolwiki) and will be accompanied by a schema that can be adapted to many different uses.

Metadata

To facilitate description of datasets and data records, metadata standards are required. Some controlled vocabularies exist, such as those provided by the Global Change Metadata Standard (GCMD), ISO 19115 and the Federal Geographic Data Committee (FGDC), but urgently need to be expanded for management of marine biology and ecology data. The searching of metadata is improved by knowing the relationships between words, such as if a word naming a concept is equal to, a subset (or child) of, or related to another word in some other way. This field of informatics, 'ontology', is well established in information science and used by librarians, but is rarely familiar to marine biologists and ecologists. Ontologies include dictionaries, controlled vocabularies, thesauri and classifications. Classifications can indicate taxonomic phylogenies and relationships between habitats and place names, and may or may not be hierarchical. They aid capture of information from the literature, as well as datasets, and are the mechanism for creating a 'semantic web' (www.semanticweb.org). However, their construction requires collaboration between ontology and marine biodiversity 'domain' experts. Such collaboration is being facilitated by the Marine Metadata Initiative (MMI).

NOMENCLATURE

In contrast to established physical ocean and genetic data management, the common element in all parts of biodiversity informatics is species names. The application of some species names changes over time, such as when a species is discovered to contain several species, or to have been described under different names. This 'concept synonymy' is a problem for information management from initial data discovery to its interpretation. TDWG is developing a 'Taxon Concept Schema' to facilitate exchange of taxonomic information, which will complement the Darwin Core and ABCD schemas designed for specimen and observation data. Although this problem is recognised (e.g. Berendsohn 1995, Geoffroy & Berendsohn 2003), it is tedious to address, because there are many more names than species, and expert knowledge is required of the history of use of each species name.

The Linnaean system of species nomenclature is the best available with well-developed rules, although codes for species names and common names are sometimes seen to have supplementary value (Froese 1999). Indeed, most users of online search engines such as FishBase and OBIS, even scientists, tend to use common instead of Latin names where these are available (e.g. Boden & Teugels 2004).

Similarly, place names change over time, and the same names may be used for different locations. Available gazetteers may find locations of some marine place names, but they do not yet intelligently link these locations to databases to integrate data. Ecological nomenclatures are also complex, with terminology for habitats and what defines ecosystems varying significantly.

Informatics should reduce duplication errors by making species names and descriptions more readily available online. Having an online register of all species names, as initiated by Species 2000 (www.species2000.org), may soon become a reality (Polaszek et al. 2005), enabling more rapid identification and avoiding the re-description of species (Costello et al. 2006, in this Theme Section). The first step towards this, having a checklist of all described species is well-underway by initiatives such as Species 2000, Integrated Taxonomic Information System (ITIS; www.itis.usda.gov), Fauna Europaea (www.faunaeur.org) and the European Register of Marine Species (www.marbef.org/data/erms.php).

The Global Biodiversity Information Facility's (GBIF) 'Electronic Catalogue of Names of Known Organisms' (ECAT) includes the Catalogue of Life (CoL), a joint publication by Species 2000 and ITIS, whose marine taxa are promoted by OBIS. CoL has listed about one-third of the estimated 1.75 million described species (Bisby et al. 2005). A parallel initiative, uBio, was founded in the library community (www.ubio.org). It is capturing all used species names from the literature to form a 'NameBank' and relating this to higher taxa in a 'ClassificationBank'. This will facilitate the location of information in libraries and online sources and, linked with the currently valid names in CoL, will greatly aid access to biological information. GBIF, with OBIS as its major partner for marine species, will use all available names (e.g. from source datasets) and, where possible, match these against the validated names in CoL.

DATA SYSTEM DESIGN

Centralised databases

The first informatics approaches to biodiversity data management were single centralised databases. These have the advantages of a single data structure and nomenclature, and are the best approach when the data are largely required within a host institution and that host is willing to undertake its management. Examples of such marine databases based on world taxa are FishBase (Froese & Pauly 2000, Boden & Teugels 2004), AlgaeBase on seaweeds and other algae (Nic Donncha & Guiry 2002), Hexacorallia on sea anemones and related taxa (Fautin 2000), CephBase on squid, octopus and related taxa (Wood et al. 2000), NeMys on mysid crustaceans and free-living nematodes (Deprez et al. 2004) and ITIS. There are several regional marine databases, for example MEDIFAUNE on Mediterranean fauna (http://nephi.unice.fr/Medifaune), MedOBIS on Mediterranean and Black Sea species (Arvantidis et al. 2006, in this Theme Section), BioOcean on deep-sea species (Fabri et al. in this Theme Section) and MASDEA on species of eastern Africa (Fondo et al. 2005, Vanden Berghe 2005), as well as a global online database based on a marine habitat, namely SeamountsOnline (Stocks 2004). However, when a database becomes larger and requires many participants, then centralised systems place a heavy technical, scientific and financial burden on a single organisation (Merali & Giles 2005). A centralised database may allow online access to the scientists who maintain the data, while the host institute focuses on technical aspects of data management; this model is in use by the European Register of Marine Species (Costello 2000, 2004, Costello et al. 2006, in this Theme Section).

Networked databases

Some recent biodiversity informatics initiatives, such as Species 2000 (Bisby et al. 2005), OBIS (Grassle & Stocks 1999, Zhang & Grassle 2003, Costello et al. 2005a), MaNIS (Stein & Wieczorek 2004) and GBIF (Edwards et al. 2000), are federations of databases distributed in many organisations around the world that agree to share data using common schema and protocols. OBIS is the data-integration component of the Census of Marine Life (CoML) (Yarnick & O'Dor 2005; www.coml.org); it will thus publish both data held in databases sourced from the literature, specimen collections and field observations, as well as new data collected by CoML field projects that address taxonomic and geographic gaps in information.

Distributed data systems have financial, quality control, ownership and community building advantages over centralised structures. The funding costs are distributed, data remain dynamic and are maintained at source by those best qualified to update and improve them, and data ownership issues are minimised, because the custodian retains control over what data are shared. Building a scientific community to support and develop the data system is promoted, because the providers of the source data remain directly involved. The central web site or 'portal' that connects all the datasets can thus concentrate on portal function rather than raw data collection and management. The costs of hardware, software and expertise are similarly distributed, and know-how can be shared amongst the participants.

There are challenges to a purely distributed system, in that the speed of response can decrease with network growth; the availability of the potential data is variable, as some sources may be off-line; the data quality varies between sources; metadata needs to evolve in parallel; the portal is ignorant of the data content, so it cannot develop advanced data handling and search tools; and users may get no feedback as to why 'zero' returns occur (this may be a case of no data or temporarily no data). One solution is to 'crawl' the data sources and 'cache' the data at intervals. Thus, the data can be classified and indexed, for example, geographically and/or taxonomically. The OBIS index is a subset of all cached data, and the indexing allows calculation of statistics on available data (Rees & Zhang in press). By resolving records in the cache to 1 record per geographic grid-square, it reduces data volume and allows more rapid online search and mapping. It allows 'near matches' to account for misspellings, and users can search down the taxonomic hierarchy. Because users are more aware of the data content, their searches can be customised. GBIF also uses an index to facilitate more rapid searching.

System support

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. An alliance of people and/or organisations with a shared vision provides synergy, and such leadership has greater impact on the scientific and government communities than the efforts of a few. Members of the alliance can share knowledge, know-how and resources such as software and ship-time. They can provide a mix of national and international matching funds for research projects, which benefit both individual members and the alliance as a whole.

QUALITY ASSURANCE

Quality assurance is especially challenging when all the possible uses to which data can be put cannot be predefined. The perceived value of data is dependent on the purpose to which they are put. Knowing a species occurs in the Pacific Ocean is useful at a global scale, but somebody else might want to know where in that ocean it occurs so that they can judge whether their discovery is a range extension. Thus, 'Pacific Ocean' is adequate quality for the first user, but not the second.

The completeness of a product is a function of its stated content, rather than the expectations or needs of the user. Unfortunately, naïve users may not appreciate that so little of the marine environment has been explored, that many species remain to be discovered, and that among species that have been observed only a fraction have been described or published in any format. Setting goals too high for a product may delay completion and publication, but setting interim goals that allow for step-wise publication provides a service for users and demonstrates progress. For example, a simple checklist of known species (a reasonably straightforward goal, but still incomplete for most of the world taxa) is seen to be of more value when it is the first step in a process in which it will in the long term provide the backbone for linking to synonyms, distribution data, identification information and published literature.

The early steps in quality assurance begin at the point of data collection (Chapman 2005a). This is followed by procedures to minimise additional errors that may arise from the processes of documentation, digitisation, archiving and publishing (either on paper or electronically). Because the opportunity for errors increases with the number of steps in handling the data, it is critical for raw data to be available in their basic form, as well as in synthesised forms. Present ocean biodiversity information systems may serve data from authoritative sources, but less credible sources, such as amateur websites and students' web pages, also exist. Quality assurance includes provision of adequate metadata, standardised data format (e.g. consistent placement of rows and columns in a table) and standard, pre-defined terminology. Quality control procedures include checks for missing values, scanning for impossible and anomalous values, mapping and graphing to check for outliers, and calculations to check that the number of records match expectations. Checking for outliers and irregularities needs expert intervention, to avoid removing apparently anomalous but nonetheless true data. The use of standard data schema enables the application of special software tools to biodiversity datasets, such as the DataTester developed by the Centro de Referência em Informação Ambiental (CRIA, Brazil) and available through GBIF (www.secretariat.gbif.net/datatester/index.jsp).

The best quality control comes from use of the data. This will be facilitated by the process of publication of primary data. User feedback must be encouraged, and this form of peer review should become a prerequisite for online data publication as it is for the publication of printed papers. Online informatics can save costs in printing, but the time and costs involved in editing, quality control and peer review may remain significant in the publication process (Kinne 1999). However, improved metadata standards may help address 2 of the problems in current science publications identified by Kinne (1999), namely by enabling more accurate search and retrieval of information from the 'growing mass of knowledge' and reduce 'wordiness and jargon'.

Conventional statistical analyses require presence and absence data. However, being certain of a species absence is challenging in ecology, because many observations are limited in space and time and all sampling methods are biased. For example, without the use of underwater video, the abundance of deep-sea coral reefs on the continental shelf of Europe would have remained unknown, although some reefs are 320 km^2 (Costello et al. 2005b). Thus, ecological studies often limit analyses to presence-only data. Museum collection data are also biased by specimens of rare species and exclude absence information. However, protocols to convert presence-only occurrence data into presence-absence may be possible if based on standard sampling and survey methods. Such tools could significantly increase the utility of online data, but they do require high compliance with metadata standards that have yet to be established.

Data quality indices could be developed based on evidence that steps in a standard quality assurance process were conducted. As mentioned previously, data suitability is a different issue and is dependent not on the data, but on the purpose for which it is required. An objective method for scoring data reliability has been utilised in FishBase (Froese et al. 1999).

CHALLENGES

Data access

Most data collection is paid for directly or indirectly by public funds with the intent that they ultimately benefit society through research, development and resource management. The failure to publish raw data undermines science, including the management of natural resources, by impeding independent analysis, as well as reuse and combination of different datasets. The calls by international scientific organisations such as the IOC and ICSU (International Council for Science 2004) to make data publicly available are being ignored by many scientists, and are thus being repeated at international conferences (Table 2). For example, NODCs contain less than half of the oceanographic data collected in their countries (Kohnke et al. 2005), and few of the marine papers in top journals publish their data. Scientists, funding agencies, institutions and publishers must require the publication of data in user-accessible form.

Table 2. Public statement by the 2004 conference on Ocean Biodiversity Informatics

> 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 organization, 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 recognizing 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

Science culture

The challenges facing OBI are not merely technological. Arguably, the greatest obstacle is the lack of a data publication culture in marine biology (and other sciences). Government agencies may make data available as a required public service, and some have realised the potential of the Internet and good data-management policies to make this a straightforward and low-cost process. Interoperability provides added benefits because, by using standard schema and protocols, data can be easily exchanged between different offices of an organisation, with related government organisations and with the international scientific community. However, unless required by funding agencies, there is no incentive for individual scientists to publish their raw data. Science journals generally prefer statistics and a synthesis of data, but an increasing number now allow data to be published as online appendices. These appendices could be published in a standard format for data exchange and, hence, facilitate interoperability if the publishers would agree to such standards (as those in molecular genetics have). Such standards exist and are in use by OBIS, GBIF and others. It is the expected practice in taxonomy to lodge type specimens in museums and, in genetics, to deposit sequences in GenBank, prior to publication. There should be a similar requirement by journals that ecological data be made publicly available prior to printed publication (International Council for Science 2004).

Froese et al. (2004) reviewed the concerns about, and excuses for not, making fisheries data available. They found that these concerns can be overcome through a combination of delayed data release, data aggregation, data use agreements, disclaimers, read-only access (the norm), data owner support and involvement, and crediting the source. The advantages of data publication are not only to other scientists, but in the long term to society (Table 1). In addition, the data providers receive more visibility, recognition, invitations, citations and collaborations (Froese et al. 2004). Indeed, publishing data may be better for 'marketing' a scientist or organisation than publishing papers, because it demonstrates an advanced level of data management. Proper recognition of online publication requires authors and editors to provide a comprehensive citation (i.e. author, year, title, publisher, url, date accessed), and for users to use the citation. Unfortunately, neither practice is yet routinely observed.

Interoperability

Emerging improvements in interoperability include: (1) more automated ways of merging datasets and cross-checking of nomenclatures (e.g. Froese 1997), (2) methods of having a 'Globally Unique Identifier' (GUID) for every data record that will allow detection of duplicate records, (3) expanded schema to allow more data and metadata to be exchanged and (4) new versions of data exchange protocols and middleware that are more comprehensive and easier to implement. With common data-sharing tools and increasing amounts of data in the public domain, the same data can be retrieved via several sources. This may be avoided, in part, by selective caching and transmitting of data, such as where OBIS does not serve GBIF datasets that it already has from other sources. Automated ways of recognising and excluding such duplication at the data record level are thus necessary. Metadata standards are being developed for marine habitats, including classifications and dictionaries. These also need to be developed for describing sampling methods so that users can appreciate the bias that may exist in datasets. Fisheries scientists have special catch-related data that require standards to facilitate interoperability not needed by other sorts of data.

Mapping

Desktop Geographical Information Systems (GIS) have now become standard in the marine and environmental sciences (including management), and GIS designed for operating online are being developed (Guralnick & Neufeld 2005, Halpin et al. 2006, in this Theme Section). Mapping as data points, routes (as lines) followed by satellite-tracked animals and polygons (areas) are available online, and ways of converting among these types mapping and comparing results to ocean data are improving. Online, semiautomated 'gazetteer' tools to translate between place names, points and polygons are being developed (e.g. www.biogeomancer.org) and will improve (Beaman et al. 2004).

Changing technologies

Computer technology is changing at such rapid rates that it is difficult to predict what opportunities will be available in future years, although monitoring the commercial sector is a good indication. OBI requires an entrepreneurial approach that seizes opportunities for technology transfer and sees change as an exciting opportunity rather than an impediment to development. Having a variety of choices in hardware and software platforms may seem confusing, but must be recognised as the normal market-driven approach in innovation. Resources are always limited and investments must weigh the uncertainties of more novel and progressive approaches against the certain needs of their market. Dealing with the uncertainties of future funding, what technologies and data will be available, and who will use the data for what purposes have parallels in any innovative business. Biologists may recognise this process as evolution. Materials (types of data), technological tools, products (e.g. maps, models, derived data) and customers are all likely to change. Thus, OBI initiatives must be adaptable to change and regularly review the way they operate.

User community

In parallel with advancing technology, the expectations of users change, and so will the culture of science. Initially, most users of OBI are probably scientists. This is essential because their use of the data is a key aspect of quality assurance, and their involvement will improve the functionality of the systems. It is also critical that the systems have the confidence of the scientific community, because, without that, further investment of experts' time and government funding will decline. Gradually, university and high school students, teachers and members of the public will make up greater numbers of users, but it will take time to develop awareness within this community. Most users of FishBase, the largest online marine biodiversity database, are from 'individual' (private) email addresses, with university-based users second (Boden & Teugels 2004). The most influential users (from a sustainability perspective) may be the relatively few scientists working for governments, universities and nongovernmental organisations. To attract scientific and education users, systems must have authoritative and credible content. Exciting tools may elicit a 'wow' factor and attract first-time users, but robust content is much more likely to result in repeated and long-term usage.

Data use index

While looking forward with imagination, there are lessons from history. One of the greatest advances in human communication was the invention of the printing press. It allowed mass production of information, much of it with no peer review or quality control. The size of libraries increased and, in time, edited science journals, and later peer review prior to publication, became established. Today, many universities use rankings of the citation rates of journals and papers to judge individual scientist's productivity and performance, and governments use this information when distributing research funding. We suggest that the Internet is a similar revolution in information availability.

A citation index for data accessions ('hits') from online databases may have similar consequences for encouraging online publication by indicating data use (Table 3). It is already possible to record new users, repeat users, usage over time and data downloads, from an online database. These measures of usage could be automated and made available online. Science abstracting services already track citations in printed publications, which provide an indication of new insights from the data used. However, for this to occur, online database managers must provide clear citation instructions, authors must use them and journals must list them with other references.

Table 3. Predictions for what Ocean Biodiversity Informatics may provide in the future

Science culture

1. Data sharing normal part of scientific process in marine biology
2. Data publication on-line becomes standard practice
3. Citation rankings of on-line publications
4. Recognition value on-line publication in individual's research performance

Informatics

1. On-line mapping of many species against selected environmental variables
2. On-line visualization as graphs, maps, movies and 3-D models
3. More automated data capture and integration option
4. Citation index for use of online data
5. Improved online data publication tools, including distribution and identification information as text, images, sounds
6. Automated translations between scripts and languages
7. Automated and permanent archiving of scholarly websites

Data available

1. All valid marine species names on-line and part of the 'Catalogue of Life'
2. Identification guides (descriptions and images) to all marine species on-line as part of a 'key of life'
3. Distributions of all marine species on-line
4. Search and map by marine habitats at global scales
5. Distributions of invasive species with predictions of future spread

Consequences for efficiency in science

1. Improved quality control in identification and taxonomy
2. Increased rate of species being described
3. New discoveries and understandings of role of biodiversity in ecosystems based on data
4. Rapid re-analysis of existing data in light of new data
5. Better management of fish stocks and natural resources through better understanding of ecosystem function and health
6. Real-time monitoring of environmental (e.g. satellite, *in situ* systems) and biological (e.g. from video, sensors) data

Ownership

At present there is relatively little external peer review prior to publication of material on scholarly websites, but these sites are recognised as credible because of the organisations and people who produced them. Some online information systems, such as ERMS and OBIS, have established Editorial Boards, with a similar function in quality assurance as the boards of scientific journals. In contrast to the scientists who volunteer time to edit and peer review papers for printed journals, their efforts directly benefit the scientific community, which retains ownership of the data. This avoids concerns that commercial publishers or institutions may profit from their contributions. This has been taken a step further by ERMS and Fauna Europaea (a register of about 130 000 land animal species in Europe). These online publications are owned by the Society for the Management of European Biodiversity Data (www.smebd.org), but all scientists who contribute to these initiatives are honorary life-members; the membership elects a council to manage the databases (Costello 2000, 2004).

Commercial use

The emergence of commercial enterprises that add value to data published online and already available in the public domain is to be welcomed. Once data is in the public domain it is a compliment to its sources when others, whether researchers, teachers, or commercial companies, use it for their purposes. Data restrictions for so-called 'non-commercial' purposes may be impossible to enforce, can be hard to define, and unnecessarily discourage entrepreneurial initiative. It is often difficult to distinguish between what is commercial or 'profit making' and what is not. Some government-owned science organisations are now commercial companies. Arguably, researchers profit when they use data to further their career, as do NGOs who use data to advance advocacy for their issues, consultants who compile data for Environmental Impact Statements, and companies that produce educational or ecotourism products using the data. However, society benefits in most cases, and the focus should not be on complex restrictions, but on facilitating publication and use.

Archiving

Archiving is a concern for electronic media. Tapes, diskettes, compact disks and other media could be given an ISBN number (International Serial Book

Number) and lodged in a copyright library for archiving, but the media would eventually deteriorate and the hardware (and perhaps software) to read them may become unavailable. Web pages are notoriously transient. However, the Internet Archive (www.archive.org) now routinely copies web pages and archives them, for which storage capacity is no longer a problem. They do not, on the other hand, archive data that is only accessible through search screens. Commercial search engines also cache web pages, but delete these as they are replaced. Procedures for database backup and mirror sites are now well established, so data will not be lost if hosted in such systems. Archives that are not compromised by hardware and software changes, and facilitate data re-use, are urgently required.

Internet access

At present, Internet access remains elusive to many people in developing countries due to poor infrastructure. However, it seems probable that reduced costs of hardware and services, and increased efficiency of satellite and wireless transmission systems, will overcome this obstacle. Indeed, this will open the 'knowledge economy' to all countries and may create a new wave of user demand and innovation at present dominated by developed countries.

CONCLUSION

An IOC-sponsored workshop that brought physical oceanographers, biologists and data managers together in 1996 was followed by a symposium on ocean data management in 2002 (Vanden Berghe et al. 2004; www.vliz.be/En/activ/events/cod/cod.htm). An international conference on 'Ocean Biodiversity Informatics' from 29 November to 1 December 2004 had >170 delegates from 37 countries and 70 presentations (from >100 offers of papers) (www.vliz.be/obi). 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 (Table 3). The consequences are positive and complementary for traditional subjects, such as taxonomy (Pennisi 2000, Costello et al. 2006, in this Theme Section), biogeography, ecology and resource management (Table 3). It will make data and information more rapidly accessible to more people than printed media and thus facilitate a more rapid and informed response by society to losses and changes in biodiversity. However, it requires a change in biological science culture to one of open-access to primary data, and a greater recognition of the value of such publication by the scientific community, including publishers, funding agencies and employers. This predicted change in science culture is already underway.

Acknowledgements. We thank A. Rees (CSIRO), R. Froese (FishBASE, OBIS), D. Hobern (GBIF), W. Berendsohn (BioCASe) and 4 anonymous reviewers for helpful comments, and M. Lane (GBIF) for her detailed improvements to the text and content of this paper. Discussions with the many participants of the 2004 Ocean Biodiversity Informatics conference, OBIS, GBIF, ERMS and colleagues at the authors' institutions also assisted the development of this paper and the wider field of Ocean Biodiversity Informatics. The funding and initiative of the Alfred P. Sloan Foundation in launching the Census of Marine Life and OBIS created the environment for this new field of scientific activity. This paper is a contribution to the following projects funded by the European Union research programmes: (1) ERMS, Marine Science and Technology programme, MAS3-CT97-0146; (2) MarBEF, Marine Biodiversity and Ecosystem Functioning, Network of Excellence (Contract 505446); (3) Species 2000 Europa, Contract EVR1-CT-2002-20011; (4) BioCASe, Biological Collection Access Service, Contract EVR1-CT2001-40017; and (5) EDIT, Toward the European Distributed Institute of Taxonomy, Contract 018340 Network of Excellence.

LITERATURE CITED

Arvanitidis C, Valavanis VD, Eleftheriou A, Costello MJ and 8 others (2006) MedOBIS: biogeographic information system for the eastern Mediterranean and Black Sea. Mar Ecol Prog Ser 316:225–230

Beaman R, Wieczorek J, Blum S (2004) Determining space from place for natural history collections in a distributed digital library environment. D-Lib Magazine 10(5) Available at www.dlib.org/dlib/may04/beaman/05beaman.html, accessed 2 June 2004

Berendsohn WG (1995) The concept of 'potential taxa' in databases. Taxon 44:207–212

Bisby (2000) The quiet revolution: biodiversity informatics and the internet. Science 289:2309–2312

Bisby FA, Ruggiero MA, Wilson KL, Cachuela-Palacio M, Kimani SW, Roskov YR, Soulier-Perkins A, Hertum J van (eds) (2005) Species 2000 & ITIS Catalogue of Life: 2005 annual checklist, CD-ROM. Species 2000, Reading

Boden G, Teugels GG (2004) Twelve years of FishBase: lessons learned. In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16], p 47–55

Chapman AD (2005a) Principles of data quality, ver 1.0. Report to the Global Biodiversity Information Facility, Copenhagen. Available at: www.gbif.org, accessed October 2005

Chapman AD (2005b) Principles of data quality—primary species and species occurrence data, ver 1.0. Report to the Global Biodiversity Information Facility, Copenhagen. Available at: www.gbif.org, accessed October 2005

Costello MJ (2000) Developing species information systems: the European Register of Marine Species. Oceanography 13(3):48–55

Costello MJ (2001) To know, research, manage, and conserve marine biodiversity. Oceanis 24(4):25–49

Costello MJ (2004) A new infrastructure for marine biology in Europe: marine biodiversity informatics. MARBEF Newsl 1:22–24

Costello MJ, Grassle JF, Zhang Y, Stocks K, Vanden Berghe E (2005a) Where is what, and what is where? Online mapping of marine species. MARBEF Newsl 2:20–22

Costello MJ, McCrea M, Freiwald A, Lundalv T and 6 others (2005b) Functional role of deep-sea cold-water *Lophelia* coral reefs as fish habitat in the north-eastern Atlantic. In: Freiwald A, Roberts JM (eds) Cold-water corals and ecosystems. Springer-Verlag, Berlin, p 771–805

Costello MJ, Bouchet P, Emblow CS, Legakis A (2006) European marine biodiversity inventory and taxonomic resources: state of art and gaps in knowledge. Mar Ecol Prog Ser 316:257–268

Deprez T, Vanden Berghe E, Vincx M (2004) NeMys: a multidisciplinary biological information system. In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16], p 57–63

Edwards JL, Lane MA, Nielsen ES (2000) Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289:2312–2314

Fabri MC, Galeron J, Larour M, Maudire G (2006) Combining the Biocean database for deep-sea benthic data with the online Ocean Biogeographic Information System. Mar Ecol Prog Ser 316:215–224

Fautin DG (2000) Electronic atlas of sea anemones: an OBIS project. Oceanography 13:66–69

Fautin DG, Fippinger P (2005) Organism occurrences in an ocean observing system. In: Proc MTS/IEEE Oceans 2005: conference and exhibition, CD publication, Washington, DC, ISBN 0-933957-33-5

Fondo EN, Osore MK, Vanden Berghe E (2004) The marine species database for eastern Africa (MASDEA). In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16], p 65–70

Frank KT, Petrie B, Choi JS, Leggett WC (2005) Trophic cascades in a formerly cod-dominated ecosystem. Science 308:1621–1623

Froese R (1997) An algorithm for identifying misspellings and synonyms in lists of scientific names of fishes. Cybium 1(3):265–280

Froese R (1999) The good, the bad, and the ugly: a critical look at species and their institutions from a user's perspective. Rev Fish Biol Fish 9:375–378

Froese R, Pauly D (eds) (2000) FishBase 2000: concepts, design and data sources, Vol 17. ICLARM Contribution 1594, ICLARM, Los Baños, Laguna

Froese R, Bailly N, Coronado GU, Pruvost P, Reyes R, Hureau JC (1999) A new procedure to evaluate fish collection databases. In: Séret B, Sire JY (eds) Proc 5th Indo-Pacific fisheries conference. Soc Fr Ichthyol, Paris, p 697–705

Froese R, Lloris D, Opitz S (2004) The need to make scientific data publicly available — concerns and possible solutions. In: Palomares MLD, Samb B, Diouf T, Vakily JM, Pauly D (eds) Fish biodiversity: local studies as basis for global inferences. Fisheries Research Report 14, ACP-EU, Brussels, p 268-271

Geoffroy M, Berendsohn WG (2003) The concept problem in taxonomy: importance, components, approaches. Schriftenr Vegetationskd 39:5–14

Grassle JF (2000) The Ocean Biogeographic Information System (OBIS): an online, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography 13:5–7

Grassle JF, Stocks KI (1999) A global Ocean Biogeographic Information System (OBIS) for the census of marine life. Oceanography 12:12–14

Guinotte JM, Bartley JD, Iqbal A, Fautin DG, Buddemeier RW (2006) Modeling habitat distribution from organism occurrences and environmental data: case study using anemonefishes and their sea anemone hosts. Mar Ecol Prog Ser 316:269–283

Guralnick R, Neufeld D (2005) Challenges building online GIS services to support global biodiversity mapping and analysis: lessons from the mountain and plains database and informatics project. Biodiversity Informatics 2:56–59

Halpin PN, Read AJ, Best BD, Hyrenbach KD and 5 others (2006) OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. Mar Ecol Prog Ser 316:239–246

International Council for Science (2004) ICSU report of the CSPR assessment panel on scientific data and information. ICSU, Paris

Jackson JBC, Kirby MX, Berger WH, Bjorndal KA and 15 others (2001) Historical over fishing and the recent collapse of coastal ecosystems. Science 293:629–638

Kaschner K, Watson R, Trites AW, Pauly D (2006) Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model. Mar Ecol Prog Ser 316:285–310

Kinne O (1999) Electronic publishing in science: changes and risks. Mar Ecol Prog Ser 180:1–5

Kohnke D, Costello MJ, Crease J, Folack J, Martinez Guingla R, Michida Y (2005) Review of the International Oceanographic Data and Information Exchange (IODE). Report submitted to the Intergovernmental Oceanographic Commission (IOC) of UNESCO, 23rd session of the assembly. Available at http://ioc3.unesco.org/iode/files.php?action=viewfile&fid=501&fcat_id=124

Lleonart J, Taconet M, Lamboeuf M (2006) Integrating information on marine species identification for fishery purposes. Mar Ecol Prog Ser 316:231–238

Lotze HK, Milewski I (2004) Two centuries of multiple human impacts and successive changes in a North Atlantic food web. Ecol Appl 14:1428–1447

Merali Z, Giles J (2005) Databases in peril. Nature 435:1010–1011

Millard K (2004) MarineXML — using XML technology for marine data interoperability. In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16], p 163–175

Myers RA (2000) The synthesis of dynamic and historical data on marine populations and communities; putting dynamics into the Ocean Biogeographic Information System (OBIS). Oceanography 13:56–59

Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. Nature 423:280–283

Nic Donnacha E, Guiry MD (2002) AlgaeBase: documenting seaweed biodiversity in Ireland and the world. Biol Environ Proc R Ir Acad 102B:185–188

Pauly D, Alder J, Bennett E, Christensen V, Tyedmers P, Watson R (2003) The future for fisheries. Science 302: 1359–1361

Pennisi E (2000) Taxonomy revival. Science 289:2306–2308

Polaszek A, Agosit D, Alonso-Zarazaga M, Beccaloni G and 24 others (2005) A universal register for animal names. Nature 437:477

Reed G, Pissierssens P (2004) New internet developments: marine XML. In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16], p 177–185

Rees T, Zhang Y (in press) Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System. In: Vanden Berghe E, et al (eds) Proceedings 'Ocean Biodiversity Informatics'—International conference on marine biodiversity data management. VLIZ Special Publication 20, Vlaams Institut voor de Zee, Oostende

Rohde RA, Muller RA (2005) Cycles in fossil diversity. Nature 434:208–210

Stein BR, Wieczorek JR (2004) Mammals of the world: MaNIS as an example of data integration in a distributed network environment. Biodiversity Informatics 1:14–22

Stevens D, Richardson AJ, Reid PC (2006) Continuous Plankton Recorder database: evolution, current uses and future directions. Mar Ecol Prog Ser 316:247–255

Stocks KI (2004) SeamountsOnline, an online information system for seamount biology. In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16], p 77–89

Vanden Berghe E (2005) MASDEA: marine species database for eastern Africa. Indian J Mar Sci 34(1):128–135

Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) (2004) Proceedings of 'The Colour of Ocean Data' symposium. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Special Publication 16]

Wiley EO, McNyset KM, Peterson AT, Robins CR, Stewart AM (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16:120–127

Wood JW, Day CL, Lee P, O'Dor RK (2000) CephBase: testing ideas for cephalopod and other species-level databases. Oceanography 13:14–20

Yarnick K, O'Dor R (2005) The census of marine life: goals, scope and strategy. Sci Mar 69(Suppl 1):201–208

Zeller D, Froese R, Pauly D (2005) On losing and recovering fisheries and marine science data. Mar Policy 29:69–73

Zhang Y, Grassle JF (2003) A portal for the ocean biogeographic information system. Oceanol Acta 25:193–197

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: December 2, 2005; Accepted: February 21, 2006
Proofs received from author(s): May 5, 2006

Vol. 316: 215–224, 2006 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published July 3

Combining the Biocean database for deep-sea benthic data with the online Ocean Biogeographic Information System

Marie-Claire Fabri[1,*], Joëlle Galéron[1], Michel Larour[2], Gilbert Maudire[2]

[1]Département Etude des Ecosystèmes Profonds, Ifremer Brest, BP 70, 29280 Plouzané, France
[2]Département Informatique et Données Marines, Ifremer Brest, BP 70, 29280 Plouzané, France

ABSTRACT: The Biocean database was designed to collate the extremely large volume of data collected from different deep-sea ecosystem studies conducted by Ifremer's department of 'Environnement Profond' (Deep-Sea Environment). This database comes in a 6-application package: 2 of them are used on research vessels to collect operational data, whereas the others are used to link with a core database back on land. The latter are used to: (1) manage taxonomic nomenclature, (2) monitor the identification of faunal collections, (3) fill in chemical analyses results or measurement data files and (4) add or extract data from the database. Biocean was designed to facilitate ecosystem studies in the deep sea. It represents an important new resource for deep-sea ecologists and will have wide applications in biogeography and biodiversity studies at Ifremer, but also for the international community, as faunal data are linked to the Census of Marine Life information system OBIS (Ocean Biogeographic Information System). Biogeographical analyses of Biocean data recovered through the OBIS portal evidence an evolution in the deep-sea sampling strategies more focused in limited areas and is intended to reveal ecosystems functioning.

KEY WORDS: Data management · Deep-sea ecosystems · Biogeography · Taxonomy · Benthos · Biodiversity · Environmental parameters

INTRODUCTION

The deep-sea environment covers 65 % of the Earth's surface and remains, for the most part, unexplored. It is characterised by very low temperature, extraordinarily high pressure and low food availability. Despite such harsh environmental conditions, the deep sea harbours generally very high levels of biodiversity (Bouchet 2000). Extrapolations from quantitative samples taken on the North Atlantic Slope and Rise suggest that there may be up to 10 million deep-sea species (Grassle & Maciolek 1992). In contrast, only 274 000 species have been taxonomically described from the entire marine environment (Bouchet et al. 2002). The discovery of hydrothermal vents in 1977, off the Galapagos Islands (Lonsdale 1977), and the discovery of cold-seep organisms on passive margins (Paull et al. 1984) have brought new knowledge of the deep sea, with these ecosystems characterised by high biomass and relatively low diversity.

Over the past 30 yr, huge numbers of samples have been collected on deep-sea cruises dedicated to benthic community studies. These faunal samples are usually sieved through a series of mesh sizes and separated into taxonomic groups at the phylum, class, or order level before being dispatched to taxonomists for identification at the most precise taxonomic level. The occurrence of species that are new to science is very frequent. Species-level identification and taxonomic description may take 5 to 10 yr after a cruise. The Biocean database was created to help manage the several steps between sampling and species identification.

The deep-sea environment can also be characterised by very strong habitat gradients and close links be-

[*]Email: marie.claire.fabri@ifremer.fr

tween species distribution and physical or chemical parameters. For example, chemosynthetic communities (such as those found at hydrothermal vents or cold seeps) live in fragmented ecosystems distributed either along ocean ridges or on active or passive margins (Sibuet & Olu 1998, Desbruyères et al. 2000). The environmental factors characterising these ecosystems represent a wide range of data, which can be stored in the Biocean database alongside faunal data.

The study of isolated communities in the deep ocean requires the use of research submersibles or remotely operated vehicles (ROVs) that are used, among other things, to capture digital pictures of *in situ* communities. These pictures represent an important information source that can be coupled to faunal identification, physical and chemical factors, as well as geographical positions. These images include crucial ecological information that is useful for habitat description, or to study species distribution and interspecific relationships.

All these data are central for ecological studies and constitute a unique resource that must be organised and permanently archived. The goals of the Biocean database are: (1) collection of operational data from research cruises, (2) organisation of faunal and environmental data in a standardised form and (3) preservation of data for studies of long-term temporal changes.

An additional goal of the Biocean project is to link with other meta-databases that collect biological information at larger scales, for example, those dedicated to biodiversity assessment (e.g. the Ocean Biogeographic Information System [OBIS]; Grassle & Stocks 1999, Grassle 2000) or global networks such as the Global Biological Information System (Edwards et al. 2000).

This paper describes in detail the development and uses of the Biocean database. Examples are provided of how Biocean data can be retrieved through the OBIS web portal and used for biogeographical studies.

MATERIALS AND METHODS

State of the art. At Ifremer's department of 'Environnement Profond', the first step towards a synthetic storage approach for biological data was the creation of a 15-table database designed in 1984. This first database was a compilation of faunal samples collected with classical oceanographic equipments, for example corers, trawls and dredges. Data collated before 1990 were compiled in this database, but neither environmental variables nor submersible data (video stills, photographs and accurate sampling positions) could be associated with faunal observations.

In 1995 a new conceptual database model was created at Ifremer. As requested by scientists, the original biological database was extended to include associated deep-sea environmental variables in a standardised form, as well as images, which have been considered an extremely valuable source of scientific information. The design and building of the Biocean database started in 1996. This crucial step was completed within a year, resulting in 56 relational tables, developed under Oracle 8i. This formed the Biocean core database.

Biocean package. The Biocean package consists of the core database, updated with a suite of applications that manage cruise and dive logs, biological samples and taxonomic names (Fig. 1).

Core database: Within the core database, tables can be considered as 1 of 3 types (Fig. 2). Metadata, for example cruise name, ship name, submersible name and dive number, are listed in the metadata tables. Reference tables list equipment, geographical locations, taxonomic specialists and faunal taxonomy, ecology and natural history observations. These data are accessed through pick-lists. Faunal, chemical and physical factors are stored in result tables. The complete relational database schema is available on the website www.ifremer.fr/isi/biocean/acces_gb/core_database_en.htm.

'Alamer' ('A la mer' or 'At sea') applications: Two 'Alamer' applications are specifically designed to collect metadata while at sea. 'Alamer' can store operations chronologically, either during a cruise or a dive. These applications incorporate pick-lists and pull-down menus that facilitate the entry and storage of key metadata in a rapid and standardised way. 'Alamer' software is designed to reference data both geographically and temporally, which is crucial for the understanding and analysis of data.

'Alamer Campagne (cruise)' collects information about the cruise itself and all the operations carried out on board. Operational data and geographical coordinates can be read in the electronic logbook, and general information, e.g. geographical region, ship navigation file, scientist name, equipment name and sample description, can also be added.

The application 'Alamer Plongée (dive)' creates a chronological report for each dive. General data, such as the geographic coordinate system (map datum), sampling and measuring equipment and the name of scientists on watch, are included. It picks up geographical coordinates directly from the submersible navigation file. Submersible data are still problematic because of the lack of accuracy of absolute geographical positioning, depending on the depth and position of the studied area. To alleviate this problem, markers are disposed on the bottom and their names are used to precisely define locations. In addition, video stills can be selected, either in real time for ROVs, e.g. ROV

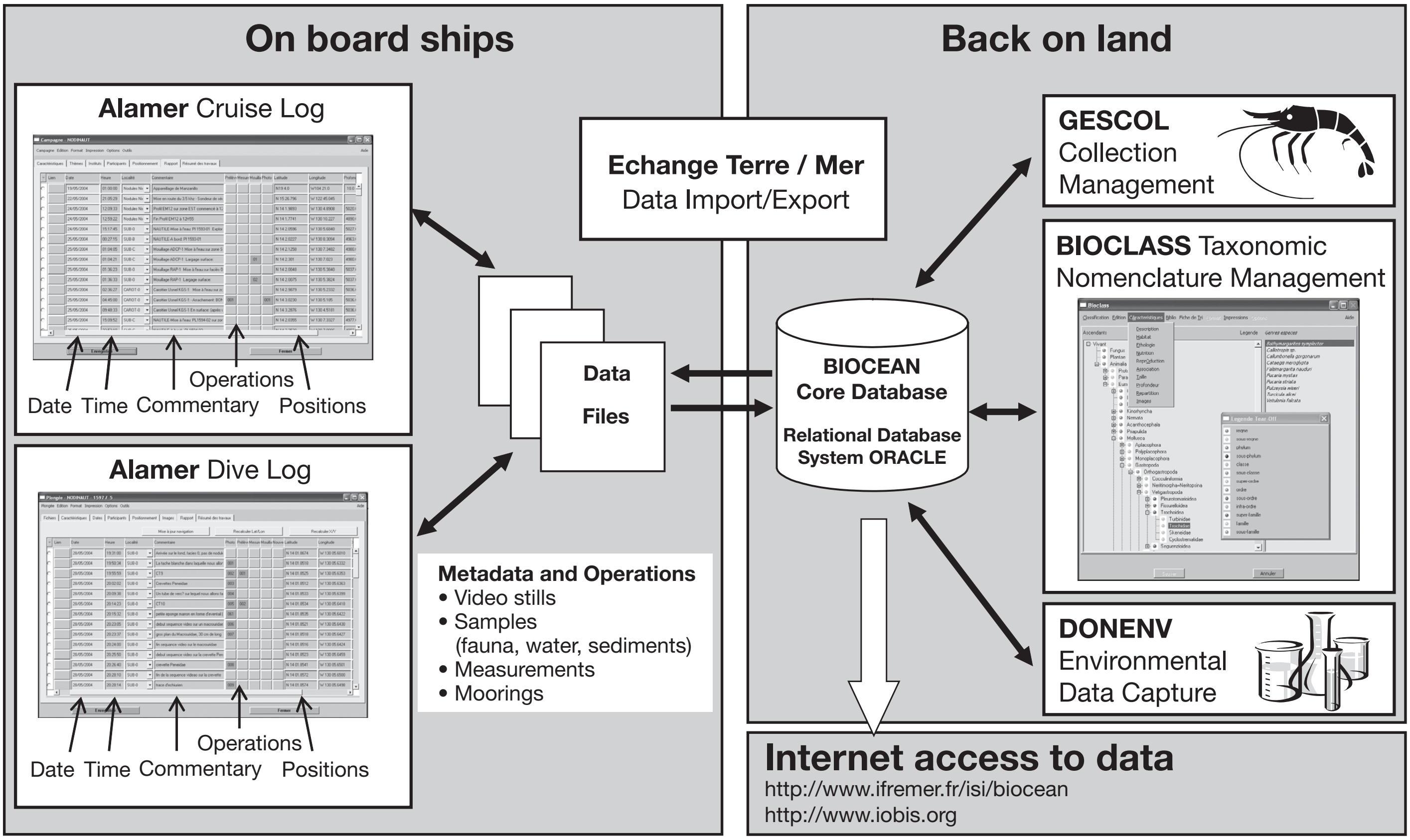


Fig. 1. Organisation of the Biocean application. The 2 Alamer applications are designed for data input directly on board research vessels, without a link to the core database. Echange Terre-Mer permits data exchange between Alamer files and the Biocean core database. Gescol follows faunal samples through their identification process. Bioclass helps to manage the taxonomic nomenclature. Donenv manages results of environmental analyses (e.g. temperature, chemistry)

'Victor', or during dive post-treatment for manned submersibles, e.g. 'Nautile'. Digital pictures are becoming essential to deep-sea ecological studies, and provide an alternative to traditional sampling methods that can potentially impact smaller fragmented habitats with low stability and resilience.

The 'Alamer' applications constitute the data input interface. The chronological reports can be enriched with event descriptions, including specific operations, such as a sediment core or water sample. Data resulting from *in situ* chemical analyses may be entered through a specific user interface, with appropriate fields for analytical method and the variable corresponding to the measured factor. Descriptions of faunal samples include names of the taxa sorted on board ship, type of preservation, sample holder name and number of individuals collected.

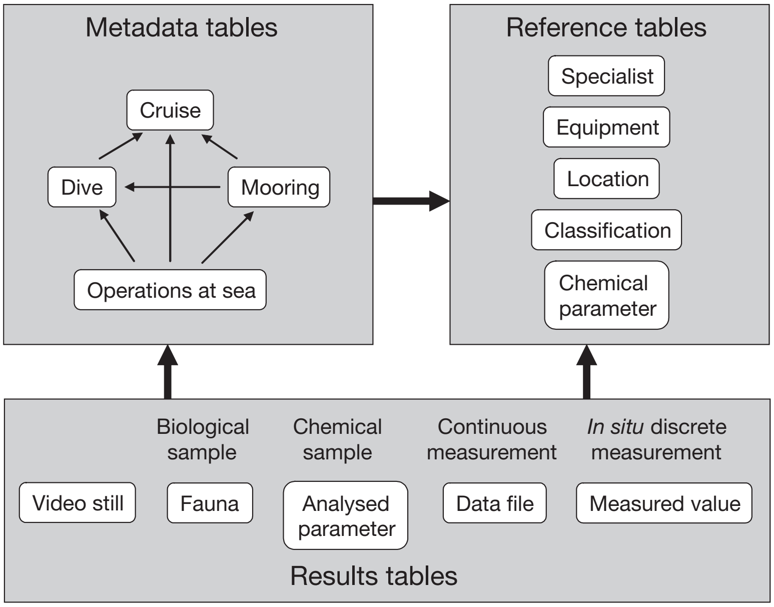


Fig. 2. Main tables of the entity relationship model of the Biocean database. See www.ifremer.fr/isi/biocean/acces_gb/core_database_en.htm for a complete relational database schema

In order to preserve the integrity of the original core Biocean database, and since more than 1 Ifremer oceanographic cruise may be occurring at the same time, all the data gathered by the Alamer applications are saved in separate files at sea. Once these data are received back on land, they can be uploaded to the core database.

'Echange Terre-Mer' application: 'Echange Terre-Mer' is the link between Alamer files and the core database. It helps to control the integrity of newly collected data before they are automatically uploaded to the database. It is used exclusively by the database administrator. This application can also be used to extract data in files for visualisation in Alamer applications.

'Bioclass' application: 'Bioclass' is a taxonomic management system that assigns codes to species or higher taxa (family, order, class, phylum). It acts as a file management system and provides a hierarchical representation of taxa. Faunal records are accompanied by a reference to the original species description and by a variety of key words related to its biology, e.g. habitat, nutrition, behaviour, reproduction and distribution.

'Gescol' application: 'Gescol' is a collection management system used to keep track of faunal samples from their collection to their identification. At sea, and in the laboratory immediately after the cruise, samples are initially divided into higher taxonomic groups (phylum, class, order). After this, they are dispatched to taxonomists and sequentially identified to the most precise level, i.e. specific level when possible. The Biocean core database is regularly updated until an accurate faunal description is obtained. 'Gescol' is also used to maintain and update a contact list of taxonomists working on deep-sea specimens.

'Donenv' application: 'Donenv' is dedicated to the data capture of chemical analysis results and data files corresponding to water or particle samples and physical measurements. These environmental parameters are stored alongside the corresponding faunal data.

Biocean products. ***'Alamer' reports:*** Individual cruise and dive reports are generated on ship, directly from 'Alamer'. They are used to share information among cruise participants and to convey initial results to scientists and institutions involved in the cruise. In addition, 'Alamer' reports include a list of operations, as well as samples and measurements, that are easily incorporated into geographical information systems (GIS). The video stills and photographs captured by submersibles are readily available for inclusion in the electronic chronological dive reports. 'Alamer' reports can provide complete documentation for each particular cruise; they are available from the website www.ifremer.fr/isi/biocean.

Core database reports: Data collected on past cruises are accessible from the core database, and may be used for ecological studies, or as references in cruise planning, when lists of moorings, markers and previous operations are required. The geographical distribution of a taxon may be retrieved from the database in tabular form, with information about the cruise (equipment, date, depth, position) and the number of individuals collected. This may be restricted to a particular station, to a specific oceanic region, or to a defined geographical area. A list of the taxa collected at a specific location can be obtained, and may be used to compare benthic faunal associations from different geographical areas. Measured physical and chemical factors can also be extracted from the database. For example, specific queries could include a list of oxygen measurements at a particular location. The mix of physical and chemical variables alongside faunal lists at a particular geographical location facilitates interdisciplinary studies that aim to understand the functioning of ecosystems.

Collection management: The 'Gescol' collection management application facilitates the generation of reports as to the whereabouts of faunal samples, and allows continuous updates regarding the level of taxonomic information (for example, from a collected new species to a published description).

Ecological data: Ecological data is managed through the 'Bioclass' application. Habitat, life-history, feeding ecology, behaviour, size ranges, associations, depth, geographical distribution, morphology (drawings and images) and taxa description can be edited and uploaded to the core database.

Technical design. The Biocean database was designed to gather together all data related to deep-sea ecology from Ifremer cruises: biological, physical and chemical. Because digital images are considered scientific data, they are stored together with the rest.

The fundamental objective that influenced the entire design of the database was an attempt to catalogue all the biological samples collected on each cruise; the process includes recording metadata on samples as soon as they are collated and recording the fate of each individual in the sample until its final identification many years later. According to the same scheme, metadata on environmental parameters are collated on board the ship and scientific data are filled in after being analysed.

These requirements have led to separate functions according to the parties involved; bilingual 'Alamer' applications are used on board by scientists from any country for the metadata collection, faunal identification and taxonomic nomenclature are followed by biologists using 'Gescol' and 'Bioclass', chemists and physicists enter water analyses and measured parameters into 'Donenv' and the database administrator's tool for core database manipulations is 'Echange Terre-Mer'.

'Alamer' applications are user-friendly; they provide menus extracted from the core database and adapted to the cruise objectives, e.g. location lists, mooring lists, or lists of faunal groups likely to be encountered, recovered, or recognised. These menus, prepared before each cruise, allow scientists to benefit from the knowledge acquired from previous cruises at the same location.

The Biocean package is considered a multimedia tool, as video stills and digital images are used as scientific data, particularly when a submersible is exploring the deep sea. The multimedia function is also used for species description in the 'Bioclass' application.

Hardware configuration. The 6 applications within the Biocean package are designed to run on personal computers under a Windows interface. Database access is possible through ODBC links. The Biocean database is held on a Unix machine under the Relational Database System (RDBS) Oracle. It is secured by an overall backup every week on magnetic cartridges (DLT), with copies conserved in separate buildings.

Recovering data. Biogeographical analyses were performed on all the Biocean data recovered through the OBIS portal. Scientific name, latitude, longitude, maximum depth and citation fields were used. Information on cruises is held in the citation field. The initial matrix was composed of 3136 species and 65 cruises. As species occurrence in the data matrix depends on the sampling effort; data were transformed into presence or absence of each species on each cruise.

A preliminary dendrogram for hierarchical clustering of all cruises, using group-average linking, based on the Bray-Curtis similarity matrix, isolated a group of 14 cruises the only shared feature of which was a very low number of species (<15). These were either recent cruises for which taxonomists had not yet identified the species, or cruises during which few specimens were collected and were omitted from the following analyses. The working matrix was composed of 3136 species and 51 cruises.

Statistical analyses were performed using PRIMER (Plymouth Routines in Multivariate Ecological Research, Ver. 5, Clarke & Warwick 2001).

RESULTS

Database content

The Biocean database contains data covering 30 yr of French deep-sea oceanographic research and includes 86 cruises, 285 dives and 442 moorings. Cruises are mainly distributed in the eastern part of the Atlantic Ocean (Fig. 3). All samples and measurements from these cruises are geographically referenced.

Metadata for each cruise are available at www.ifremer.fr/isi/biocean. They are displayed as cruise

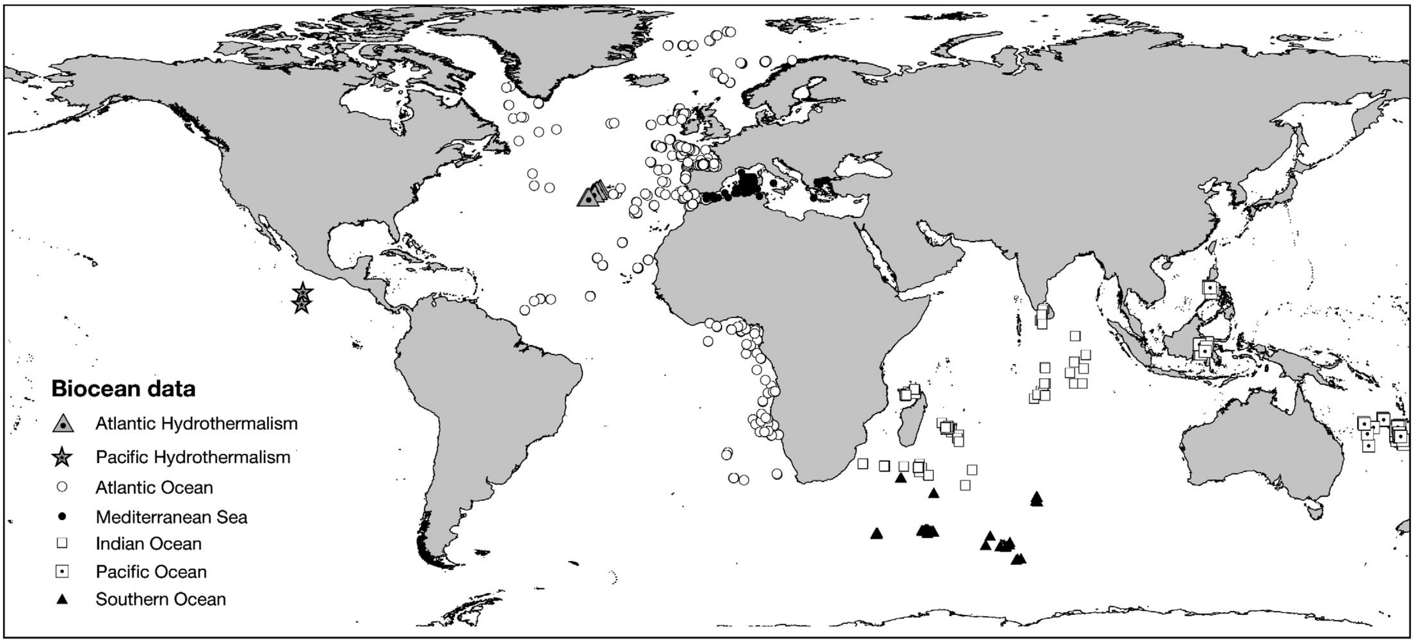


Fig. 3. Distribution map of deep-sea benthic faunal data stored in the Biocean core database and available through the Ocean Biogeographic Information System (OBIS) portal

and dive logs. Biogeographical data are available through the OBIS portal (www.iobis.org).

Database context

Ecological studies in the deep sea are focused on the structure of benthic communities and their spatial and temporal variations. Biocean holds 4 types of ecosystem data: (1) benthic sedimentary ecosystems, which depend on energetic contributions from photosynthetic production in euphotic layers and (2) deep-sea hydrothermal ecosystems on oceanic ridges, which are based on chemosynthetic bacterial production. Life in this habitat is very unusual, luxuriant and adapted to a toxic and unstable environment. There are also: (3) ecosystems associated with cold seeps on continental margins, which are based on chemosynthetic production from methane-rich fluids. These ecosystems are characterised by high biomass. And finally Biocean contains data on: (4) deep coral reefs, which have been discovered on carbonate mounds on the Irish continental margin. Active mound genesis is due to intense coral growth.

The last 2 ecosystems have been topics of recent studies at Ifremer. Species identifications are not complete, and data are consequently not yet available through the OBIS portal.

Biogeographical analyses of Biocean data

Ordination of cruises in the working matrix

The aim of ordination was to display the biological relationships among the 51 cruises. Similarities between cruises based on species presence or absence are given by the Bray-Curtis similarity coefficients. The ordination maps cruises in 2 dimensions, along which the placement of cruises reflects the similarities of their biological communities. A non-metric, multidimensional scaling (MDS) plot was produced by PRIMER (Fig. 4). A stress value of 0.12 was derived by statistical processing. As it is close to 0.10, it indicates good ordination. A cross-check against the results of hierarchical classification was made by superimposition of the groups having 10 % similarities, corresponding to the major ocean areas. Interpretation of the cruise ordination leads to the 7 clusters that appear in the figure. The Atlantic Ocean is such a compact group, apart from the hydrothermal-vent cruises, that it is discussed separately below.

The Balgim 84 cruise took place in the Straits of Gibraltar. The Straits of Gibraltar show a species composition of either Atlantic (for epibenthic crustaceans; Abello et al. 2002) or Mediterranean affinity (for Bivalvia; Salas 1995), depending on the taxonomical

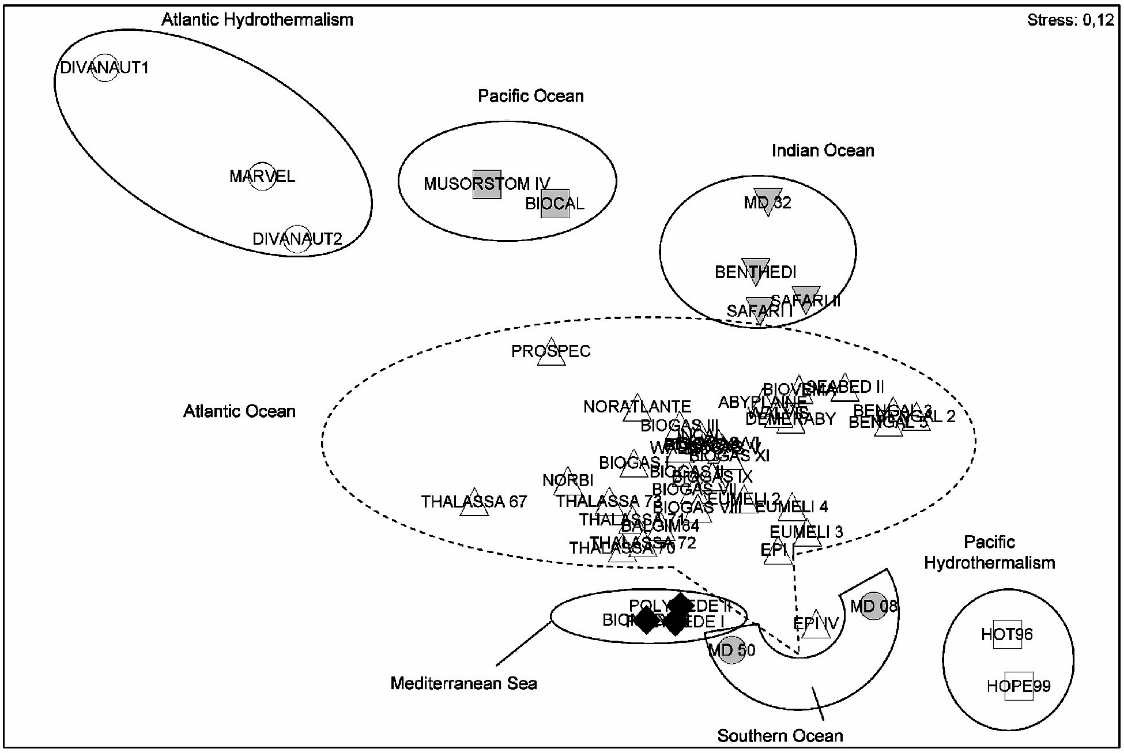


Fig. 4. Ordination plot of cruises in the working matrix. The multi-dimensional scaling (MDS) was constructed from Bray-Curtis similarity coefficients of the presence or absence of species on each cruise

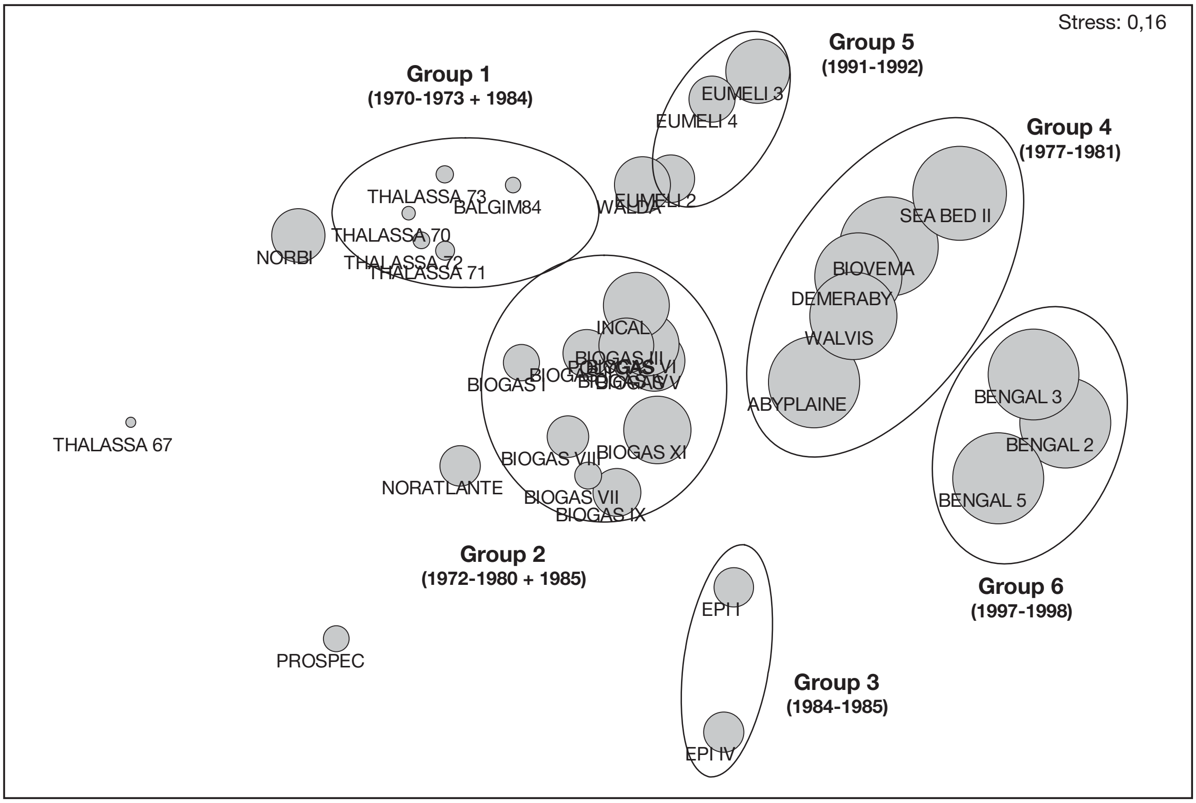


Fig. 5. MDS ordination of Atlantic cruises stored in the Biocean database. Bray-Curtis similarity coefficients of species presence or absence were used for the ordination. Clusters formed at the 15% similarity level and mean depth as circles of differing sizes were superimposed on the MDS plot

group considered. MDS ordination groups the Balgim 84 cruise and Atlantic cruises, suggesting that biological samples from the Straits of Gibraltar are composed mainly of Atlantic species.

Ordination of Atlantic cruises stored in the Biocean database

The cluster of 35 non-hydrothermal Atlantic cruises was analysed with Bray-Curtis similarity coefficients as above and MDS ordination. A stress value of 0.16 gives a potentially useful 2-dimensional picture. A cross-check against the results of hierarchical classification was made by superimposition of the groups having 15% similarities. The dendrogram suggested a division of the cruises into 6 main clusters and 5 isolated cruises. For each of the cruises, 8 abiotic variables obtained through the OBIS portal were calculated: mean depth, mean latitude, mean longitude, their standard deviations, year and number of species.

The selection of the abiotic-variable subset maximising rank correlation between biotic and abiotic similarity matrices was conducted with the BVSTEP procedure of PRIMER software (Clarke & Warwick 2001). The best 2-variable combination involved the mean depth and the year of the cruise. The Spearman rank correlation coefficient ($r = 0.65$; $p < 0.001$) gave a significant result. The most graphic representation of the result involved superimposing the mean depth on the biotic ordination as circles of differing sizes depending on the value (Fig. 5). The distribution of groups and isolated cruises in the Atlantic Ocean were mapped for better interpretation (Fig. 6).

Group 1 included cruises (1970 to 1973 and 1984) representative of the bathyal zone, with mean depths between 700 and 1000 m. Samples from these cruises were all from the top of the continental slope of the NE Atlantic, including the Straits of Gibraltar.

Group 2 included cruises carried out in the NE Atlantic during 8 yr (1972 to 1980 and 1985), at mean depths of around 2500 m and high depth variation, representing the bottom part of the continental slope.

Group 3 included the 2 Epi cruises (1984 to 1985), which were expected to cluster with Group 2 as they sampled the same location in the Bay of Biscay. The reason for their isolation was probably the very low number of species identified from these cruises; the Epi programme dealt with faunal-group densities, and few species were identified.

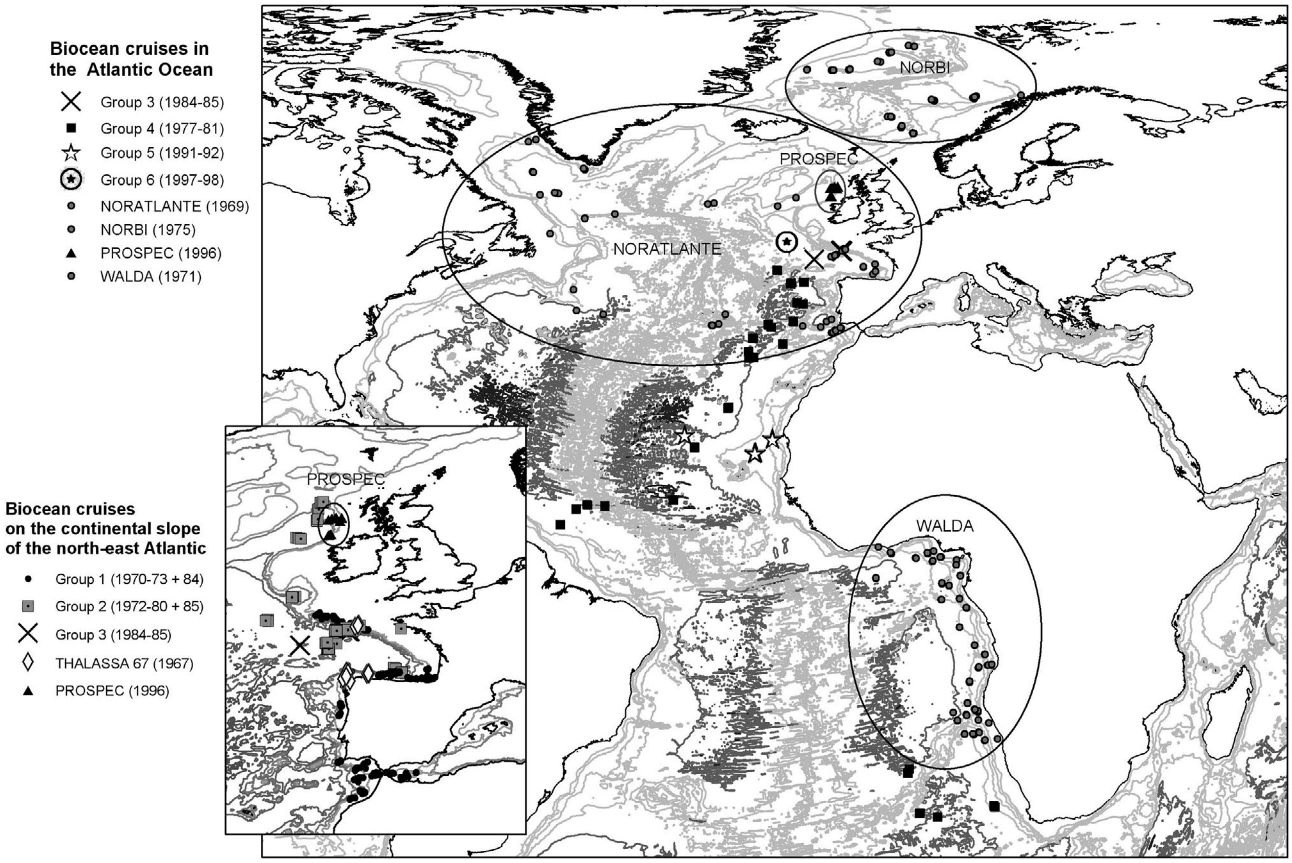


Fig. 6. Distribution map of Atlantic cruise sampling points stored in the Biocean database. Cruises are grouped with Bray-Curtis similarity coefficients based on species presence or absence

Group 4 included Atlantic cruises (1977 to 1981) that were explorations of the deep sea (4800 m), at a wide variety of latitudes and longitudes.

Group 5 included 3 Eumeli cruises (1991 to 1992) that sampled the abyssal plain in a study of the influence of particulate fluxes at 3 sites designated eutrophe (1800 m), mesotrophe (3000 m) and oligotrophe (4500 m), located near one another.

Group 6 was composed of the 3 Bengal cruises (1997 to 1998), which sampled the same point on the Porcupine Plain, characterised by a great and constant depth (4800 m). Their goal was to study seasonal effects on benthic fauna.

The 5 isolated cruises have extreme characteristics. Thalassa 67 (1967) is the shallowest cruise of the Biocean database and the oldest one. It sampled the very top part of the continental slope. Noratlante (1969) has the greatest variation in depth, as well as in latitude. It covered the eastern and western part of the North Atlantic Ocean. Walda (1971) extended along the whole west coast of South Africa, and represents the largest variation in longitude. Norbi (1975) covered in the deep Norwegian Sea, north of the threshold of Rockall. Thresholds usually create natural geographical barriers and therefore separate distinct faunal groups. Prospec (1996) is also geographically different from all of the other cruises. It studied deep-sea fishery resources on the continental slope north-west of Ireland. Species were not identified for all taxonomic groups.

DISCUSSION

Biocean package

The Biocean package and the database that is acquired, managed and archived with it were designed to optimise the organisation and accessibility of metadata, reference lists and results from shipboard sampling and shore-based analysis of marine faunas and environments. The experiences of other institutions that have developed data-management systems with similar objectives were taken into account during design. The Video Annotation and Reference System (VARS) is a software interface and database system that provides tools for describing, cataloguing, retrieving and viewing the visual, descriptive and quantitative data associated with MBARI's deep-sea video archives (www.

mbari.org/vars/vars_overview.html). The 'Alamer Plongée (dive)' and 'Bioclass' application were drawn, respectively, from the MBARI's annotation interface and the taxonomic part of MBARI's knowledge base. The Biocean multimedia function was strongly inspired by MBARI's VARS, but Biocean's fundamental objective is to track faunal samples through their successive identification steps, which is not a goal at MBARI. Applications for this purpose were influenced by the Ocean Drilling Program (ODP) Curation Janus database (http://www-odp.tamu.edu/database), which contains 450 tables of ODP's marine geoscience data that were adapted to biological deep-sea cruises and samples.

Environmental data management is an original element added to the former biological database. Chemical and physical data have long been archived in data centres such as the French National Oceanographic Data Centre SISMER (www.ifremer.fr/sismer), which designs and operates scientific information systems and databases for national and international projects in the marine domain. Environmental data management in Biocean was designed to be compatible with this national and international framework.

Databases oriented toward species distribution and taxonomy are numerous, and many are available on the Internet (e.g. CephBase at www.cephbase.utmb.edu and FishBase at www.fishbase.org). They generally deal with 1 group of taxa, whereas Biocean deals with all benthic taxa encountered in the deep sea.

Biocean differs from all other databases in that it integrates physical and chemical data with data on biology. To our knowledge, no system equivalent to Biocean exists that centralises metadata from entire cruises; that tracks samples through to their final analysis, whether taxonomic or chemical; and that can be used to retrieve faunistic data and environmental attributes.

Short-term advantages

Past decades have seen tremendous growth in data acquisition capability, arising from the use of new submersibles, such as the ROV 'Victor' 6000 and the development of improved sensors, cameras and sampling tools. The increasing amount of data collected during each cruise creates new challenges for finding solutions for the storage, access, organisation and synthesis of data.

The data management strategy of the Biocean package is to collect cruise data directly onboard using Alamer applications, so that a comprehensive and standardised dataset can be made available directly after the cruise on the website (www.ifremer.fr/isi/biocean). The applications are useful onboard to increase the efficiency of the cruise and can be used to get a precise overview of the scientific programme, as well as the data and samples collected during a specific cruise. Alamer applications can collect chronological events, print thematic lists of operations and retrieve metadata related to these operations.

Alamer applications designed to collate cruise and dive logs are available on every Ifremer ship and have been used by most of the European teams working in the deep sea. This logbook initiates a collaborative work between biologists from several fields to establish comprehensive information on a specific study area.

Long-term advantages

Data acquisition in oceanographic sciences is carried out both on cruises and during sample analyses following a cruise. Because many deep-sea habitats have only been recently discovered and harbour high biodiversity, it can take 5 to 10 yr before taxonomic data is made available for input into the database. The structure of the Biocean package allows continuous data input for many years after the cruise, and permits valuable environmental and geographical data to be linked to taxonomic information in perpetuity. The Biocean package is an essential tool to follow the ecology of a studied area with respect to aspects such as seasonal variability in abyssal plains or temporal evolution of hydrothermal vents.

Data are retrieved under tabular forms; hence, they can be easily plotted and mapped in GIS. The technological evolution of Biocean is ensured by Ifremer's Marine Technology and Information System Direction; the relational database is continuously upgraded as new versions of RDBS Oracle are released.

Internet access

This global management system is meant to archive and centralise multidisciplinary data dealing with abyssal ecology in a standardised form. To answer the increasing international demand for a census of biodiversity data, Biocean contributes to the European programme Biodiversity Collection Access Service for Europe (BioCASE 2002) and to the network Census of Marine Life (CoML; www.coml.org). Biogeographical data are available through the CoML web portal of OBIS (www.iobis.org). Metadata for each cruise are available at www.ifremer.fr/isi/biocean, as are the database model and contact details for anyone seeking scientific collaboration or requiring technical information on how to obtain the applications.

Biogeographical analyses

The bathymetric distribution of species is a concept widely expressed in local studies (Howell et al. 2002, Olabarria 2005) and obviously influenced ordination of Atlantic cruises. The ordination suggested also the influence of cruise date, which reflected evolution of sampling strategies. Older cruises, conducted during the earliest days of deep-sea oceanography, were essentially exploratory expeditions along the continental slopes. Since the 1980s an important change in the approach of deep-sea biological research has taken place. Descriptive oceanography has given way to more comprehensive studies. Sampling strategies have become more focused on limited areas in the deep sea and are intended to reveal the functioning of ecosystems.

Acknowledgements. We are grateful to D. Desbruyères, M. Sibuet and G. Riou, who supported the authors by providing the time and resources necessary to develop this application. We thank A. B. Thistle, A. G. Glover and C. L. Van Dover for help with English grammar and spelling. Many thanks to the anonymous referees, whose comments considerably improved the manuscript.

LITERATURE CITED

Abello P, Carbonell A, Torres P (2002) Biogeography of epibenthic crustaceans on the shelf and upper slope off the Iberian Peninsula Mediterranean coasts: implications for the establishment of natural management areas. Sci Mar 66:183–198

BioCASE (2002) Biodiversity Collection Access Service for Europe. Available at: http://biocase.org

Bouchet P (2000) L'insaisissable inventaire des espèces. Recherche 333:40–45

Bouchet P, Lozouet P, Maestrati P, Heros V (2002) Assessing the magnitude of species richness in tropical marine environments: exceptionally high numbers of molluscs at a New Caledonia site. Biol J Linn Soc 75:421–436

Clarke KR, Warwick RM (2001) Change in marine communities: an approach to statistical analysis and interpretation, 2nd edn. PRIMER-E, Plymouth

Desbruyères D, Almeida A, Biscoito M, Comtet T, Khripounoff A, Le Bris N, Sarradin PM, Segonzac M (2000) A review of the distribution of hydrothermal vent communities along the northern mid-Atlantic Ridge: dispersal vs. environmental controls. Hydrobiologia 440:201–216

Edwards JL, Lane MA, Nielsen ES (2000) Interoperability of biodiversity databases: biodiversity information on every desktop. Science 289:2312–2314

Grassle JF (2000) The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modeling and mapping marine biological data in a multidimensional geographic context. Oceanography 13:5–7

Grassle JF, Maciolek NJ (1992) Deep-sea species richness: regional and local diversity estimates from quantitative bottom samples. Am Nat 139:313–341

Grassle JF, Stocks KI (1999) A Global Ocean Biogeographic Information System (OBIS) for the census of marine life. Oceanography 12:12–14

Howell KL, Billet DSM, Tyler PA (2002) Depth-related distribution and abundance of seastars (Echinodermata: Asteroida) in the Porcupine Seabight and Porcupine Abyssal Plain, NE Atlantic. Deep-Sea Res I 49:1901–1920

Lonsdale PF (1977) Clustering of suspension-feeding macrobenthos near abyssal hydrothermal vents at oceanic spreading centers. Deep-Sea Res 24:857–863

Olabarria C (2005) Patterns of bathymetric zonation of bivalves in the Porcupine Seabight and adjacent abyssal plain, NE Atlantic. Deep-Sea Res I 52:15–31

Paull CK, Hecker B, Commeau R, Freeman-Lynde RP and 6 others (1984) Biological communities at the Florida escarpment resemble hydrothermal vent taxa. Science 226: 965–967

Salas C (1995) Marine bivalves from the southern Iberian Peninsula collected during the Balgim and Fauna 1 expeditions. Haliotis 25:33–100

Sibuet M, Olu K (1998) Biogeography, biodiversity and fluid dependence of deep-sea cold-seep communities at active and passive margins. Deep-Sea Res II 45:517–567

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: November 25, 2004; Accepted: December 12, 2005
Proofs received from author(s): May 10, 2006

Vol. 316: 225–230, 2006 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published July 3

MedOBIS: biogeographic information system for the eastern Mediterranean and Black Sea

C. Arvanitidis[1,*], V. D. Valavanis[1], A. Eleftheriou[1], M. J. Costello[2], S. Faulwetter[1], P. Gotsis[1], M. S. Kitsos[3], I. Kirmtzoglou[3], A. Zenetos[4], A. Petrov[5], B. Galil[6], N. Papageorgiou[1]

[1]Institute of Marine Biology and Genetics, Hellenic Centre for Marine Research, PO Box 2214, Iraklion, 71003 Crete, Greece
[2]Leigh Marine Laboratory, University of Auckland, PO Box 349, Warkworth, New Zealand
[3]School of Biology, Laboratory of Zoology, Aristotelian University of Thessaloniki, 51424 Thessaloniki, Greece
[4]Institute of Oceanography, Hellenic Centre for Marine Research, PO Box 712, Anavyssos, 19013 Attiki, Greece
[5]Institute of Biology of the Southern Seas, Sevastopol 99011, Ukraine
[6]National Institute of Oceanography, Oceanographic and Limnological Research, PO Box 8030, Haifa 31081, Israel

ABSTRACT: Recent online initiatives in sharing marine biological data, such as the European Register of Marine Species (ERMS) and the Ocean Biogeographic Information System (OBIS), identified gaps in data from the eastern Mediterranean and Black Sea. Such data are now being collected, formatted and disseminated by MedOBIS (the Mediterranean Ocean Biogeographic Information System) initiative involving Greece, the Ukraine and Israel (test version available at: www.medobis.org). The aim is to develop a taxon-based biogeography database and online data server with links to survey and satellite environmental data. MedOBIS is currently undergoing 4 stages of development, namely, data assembly, formatting, analysis and dissemination. The primary features of the MedOBIS application are its offline GIS (Geographic Information Systems) data formatting capabilities and its online Java- and JavaScript-enabling data server with taxon-based search, mapping and data downloading capabilities. It is an independent source of biological and environmental data, as well as an online GIS tool designed to facilitate access to historical and current data by marine researchers. As more data become available and are inserted into the system, MedOBIS will function as the eastern Mediterranean and Black Sea node of EurOBIS (the European node of the international OBIS initiative, part of the 'Census of Marine Life').

KEY WORDS: Marine biodiversity · Data management · GIS · OBIS

INTRODUCTION

The international and interdisciplinary nature of ecosystem research and management can be facilitated by the Internet and associated activities in biodiversity informatics. The free dissemination of valuable historical and current biological, environmental and genetic information is contributing to the establishment of an interdisciplinary research platform targeted towards information integration at regional and global scales and to the development of information-based management schemes.

Development of systems such as OBIS (Ocean Biogeographic Information System, www.iobis.org/), OBIS-SEAMAP (Spatial Ecological Analysis of Megavertebrate Populations, http://seamap.env.duke.edu/), FIGIS (FAO Fisheries Global Information System, www.fao.org/fi/figis/) and aphia (North Sea species register, www.vliz.be/vmdcdata/aphia/) facilitate the study of anthropogenic impacts on threatened species. At the same time they enhance our ability to test biogeographic and biodiversity models, support modelling efforts to predict distribution changes in response to environmental change and provide new opportuni-

[*]Email: arvanitidis@her.hcmr.gr

ties for public outreach. In addition, such online database systems allow the development of management practices that are based on synthetic analyses of interdisciplinary data (Schalk 1998, Laitinen & Neuvonen 2001, Decker & O'Dor 2002, Marshall 2002, Tsontos & Kiefer 2002, Babu 2003).

An analysis of the availability of data, expertise and identification guides concerning marine species was conducted (e.g. Costello 2000, Costello et al. 2001, 2006, in this Theme Section) as part of the European Register of Marine Species (ERMS). These activities identified the eastern Mediterranean and Black Sea as gaps in the availability of marine data. Similarly, OBIS identified the need to establish a marine biodiversity informatics infrastructure for the same region and the associated North African countries. A number of web sites provide marine biodiversity information for the Mediterranean and Black Sea region, namely, the Commission Internationale pour l'Exploration Scientifique de la Méditerranée (IESM) atlas of exotic species (www.ciesm.org/online/atlas/index.htm), a site on Mediterranean sea turtles (www.euroturtle.org/), the Mediterranean oceanic database on oceanographic research (modb.oce.ulg.ac.be/), Blackseaweb (www.blackseaweb.net/), a site on the ecology of *Mnemiopsis leidyi* (www.issg.org/database/species/ecology.asp?si=95&fr=1&sts=) and the IODE (www.iode.org). However, none of these provides information concerning the distribution and abundance of marine species in the region. To this end, the development of a new online marine biological information system called MedOBIS (Mediterranean Ocean Biogeographic Information System) was proposed.

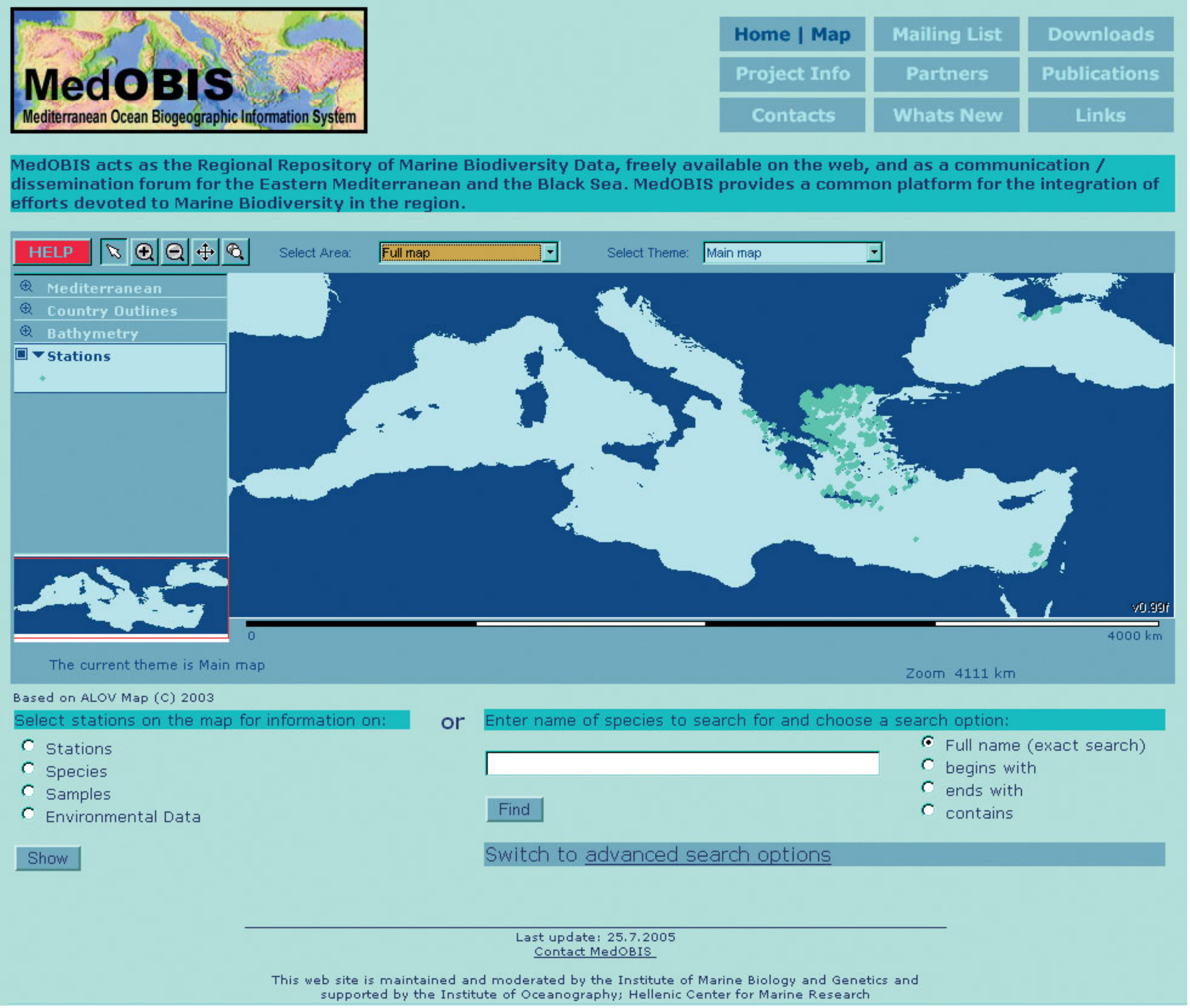
MedOBIS aims to assemble, formulate, synthesise and disseminate marine biological data for the eastern Mediterranean and Black Sea regions, focusing on the reliability and longevity of historically surveyed data, the assembly of current and new information and the dissemination of raw and integrated biological and environmental data and future products through the Internet.

DESCRIPTION OF MedOBIS

There are 4 main stages in the current development of MedOBIS: data assembly, formatting, analysis and dissemination. The data assembly phase is based on the free contribution of biological data by various national and international scientific surveys in the region, as well as the processing of time series of remotely sensed and station environmental parameters. Assembled datasets include 839 stations with surveyed benthic biological data (Fig. 1). These data consist mainly of benthic species abundance and biomass for >3000 benthic organisms, seabed substrate types and for several environmental parameters. So far datasets have been assembled from 100 stations in the Ionian Sea, 569 stations in the Aegean Sea, 106 stations in the Black Sea and 64 stations in the Levatine Sea. These data cover the period 1937 to 2003; however, most of the data relate to the period 1986 to 1996 (Table 1). In addition, a variety of satellite and station environmental data on sea surface temperature, chlorophyll *a*, photosynthetically active radiation, salinity, sea level anomaly, precipitation and wind force and direction were compiled for the period 1985 to 2004. Sources of these data include international online data

Table 1. Overview of the current taxonomic, geographic and temporal coverage of the MedOBIS database

Taxon	No. of taxa recorded	No. of observations
Taxonomic records		
Actinopterygii	45	751
Annelida	1093	9810
Anthozoa	101	440
Ascidiacea	54	350
Brachiopoda	4	6
Bryozoa	67	255
Cephalochordata	1	20
Chaetognatha	1	1
Chironomidae	4	31
Crustacea	1115	7930
Echinodermata	118	954
Echiura	2	54
Elasmobranchii	3	43
Enteropneusta	3	6
Foraminifera	11	25
Gastrotricha	1	4
Hydrozoa	10	38
Mollusca	1381	5400
Nematoda	2	39
Nemertea	13	187
Phoronida	5	100
Platyhelminthes	10	20
Pogonophora	1	1
Porifera	128	883
Priapulida	1	3
Protozoa	1	1
Pycnogonida	6	34
Scyphozoa	2	32
Sipuncula	18	271
Total	**4201**	**27689**
Geographic coverage		
	No. of stations surveyed	
Ionian Sea	100	
Aegean Sea	569	
Black Sea	106	
Levantine Sea	64	
Temporal coverage		
Years	No. of stations surveyed	
1979–1980	167	
1991–2000	516	
2001–present	128	
1937–2003 (total temporal coverage)		839



MedOBIS acts as the Regional Repository of Marine Biodiversity Data, freely available on the web, and as a communication / dissemination forum for the Eastern Mediterranean and the Black Sea. MedOBIS provides a common platform for the integration of efforts devoted to Marine Biodiversity in the region.

Select Area: Full map

Select Theme: Main map

The current theme is Main map

Zoom 4111 km

Based on ALOV Map (C) 2003

Select stations on the map for information on:

○ Stations
○ Species
○ Samples
○ Environmental Data

Enter name of species to search for and choose a search option:

● Full name (exact search)
○ begins with
○ ends with
○ contains

Switch to advanced search options

Last update: 25.7.2005
Contact MedOBIS

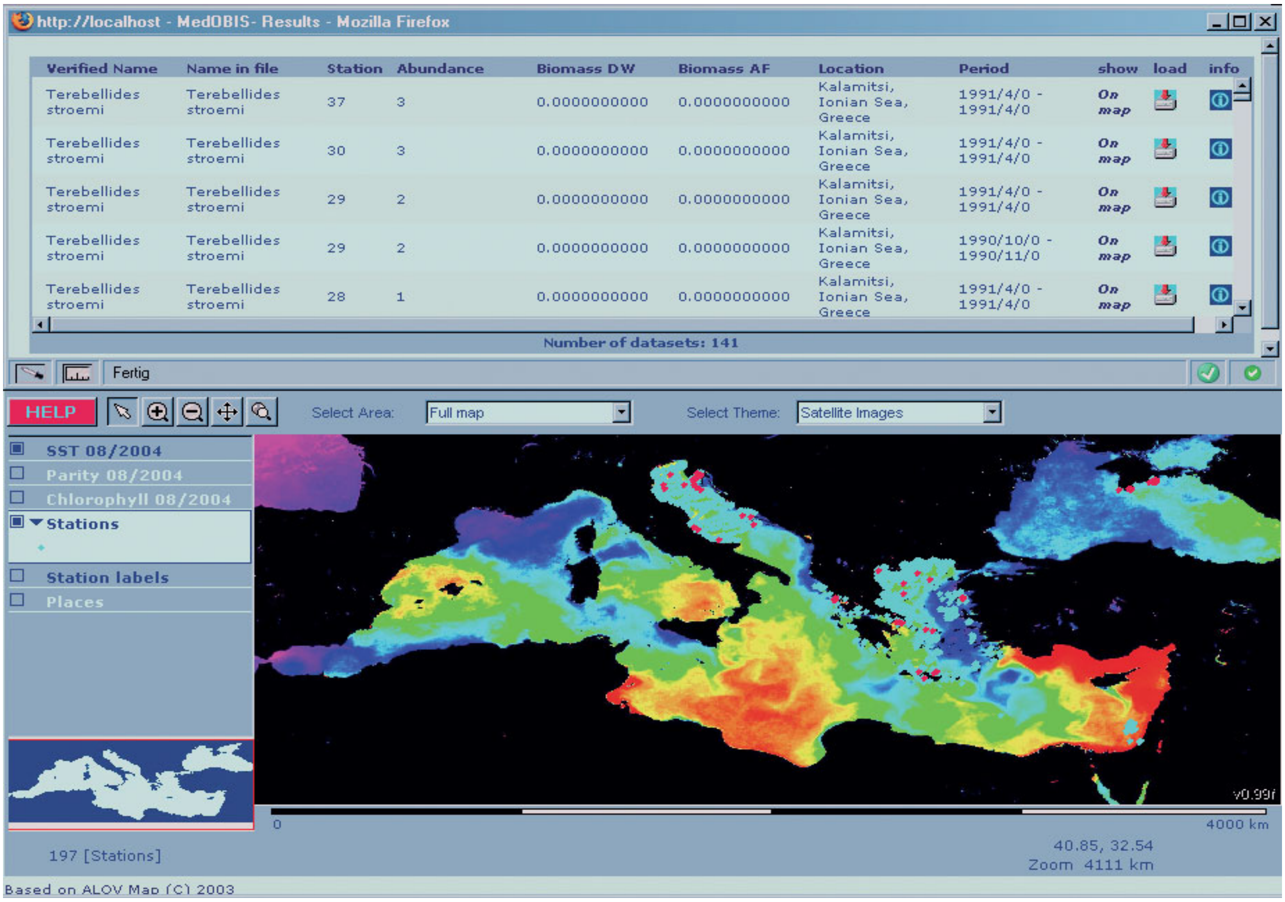
This web site is maintained and moderated by the Institute of Marine Biology and Genetics and supported by the Institute of Oceanography; Hellenic Center for Marine Research

Fig. 1. MedOBIS system interface: spatial distribution of the current assemblage of biological survey stations in the eastern Mediterranean and Black Sea (total number of stations = 839). The figure quality corresponds to that of the computer screen image

archives, such as the US National Aeronautics and Space Administration (NASA), the German Aerospace Agency (DLR) and the French European Remote Sensing (ERS) Processing and Archiving Facility (CERSAT). Currently, all assembled biological datasets are representative of the benthic communities existing in the eastern Mediterranean and Black Sea, while environmental datasets are representative of the entire Mediterranean Sea and the Black Sea. The long-term aim of the MedOBIS data assembly phase is to collect sufficient data to fully represent the changes and time intervals associated with the biological characteristics of the region. Since biological data from the eastern Mediterranean and Black Sea region are scarce in the relevant literature or, in many cases, remain undocumented, MedOBIS contributes a new geographic area to OBIS.

The data formatting stage of MedOBIS is based on a Geographical Information System (GIS) (ESRI 1994), under which all assembled datasets receive quality control and are processed under a common georeference scheme. The current quality control mechanisms include (1) the nomenclature control, based on readily available services on the web (ERMS, Species 2000 project) and on taxonomists' expertise; (2) the distribution of each species in the eastern Mediterranean and the Black Sea region, based on specialists' knowledge; and (3) the storage of the verified name in an extra attribute field — the initial name is kept to document any further changes. At a later stage, data quality control procedures will be formulated which will encompass database integrity constraints, automated plausibility checks and the synchronisation of species lists between the EurOBIS and MedOBIS databases. The GIS shapefile is the standard file format for linking the geographic coordinates of a point of interest with attributes such as measurements and specific characteristics. However, it was the complexity of the collected data that dictated the choice of the relational database management system (RDBMS) backend to store them. In this way, the GIS information layer includes the geographic coordinates of the stations as well as the station's identification number, which refers to relevant records in a MySQL database. The database fields are implemented according to the OBIS Schema 1.0, so that they will be accessible via the main OBIS portal through the DiGIR protocol (http://digir.sourceforge.net/) at a later stage of the project. Finally, satellite data are embedded in a GIS database as GIS regular grids (Valavanis et al. 1998, Valavanis 2002). The data formatting phase aims to contribute to the data interoperability issue through the production of commonly formatted GIS-ready data layers from currently scattered datasets stored in various and different formats; this is the first time such a task has been undertaken using biological data from the eastern Mediterranean and the Black Sea.

The MedOBIS data analysis phase is based on raster and vector datasets, integration through GIS and spatial analyses for the production of species distribution maps and identification of species–environment relations. Although still in its initial stage and currently only functional offline the data analysis phase has already produced several analytical results, especially on the mapping of ocean production processes and the biogeography of benthic polychaetes in the region (Arvanitidis et al. 2002, Valavanis et al. 2004a,b, in press). The data analysis phase aims to enhance the overall functionality of the system by providing a variety of spatial query tools for visualising relationships among species and their environment. Mapping of certain oceanic processes, such as marine productivity hotspots, mesoscale thermal fronts and productive gyres, will be introduced into the system as separate data layers. Thus, the user will be able to identify relationships of selected taxa to environmental gradients and processes (Fig. 2).



http://localhost - MedOBIS- Results - Mozilla Firefox

Verified Name	Name in file	Station	Abundance	Biomass DW	Biomass AF	Location	Period	show	load	info
Terebellides stroemi	Terebellides stroemi	37	3	0.0000000000	0.0000000000	Kalamitsi, Ionian Sea, Greece	1991/4/0 - 1991/4/0	On map		
Terebellides stroemi	Terebellides stroemi	30	3	0.0000000000	0.0000000000	Kalamitsi, Ionian Sea, Greece	1991/4/0 - 1991/4/0	On map		
Terebellides stroemi	Terebellides stroemi	29	2	0.0000000000	0.0000000000	Kalamitsi, Ionian Sea, Greece	1991/4/0 - 1991/4/0	On map		
Terebellides stroemi	Terebellides stroemi	29	2	0.0000000000	0.0000000000	Kalamitsi, Ionian Sea, Greece	1990/10/0 - 1990/11/0	On map		
Terebellides stroemi	Terebellides stroemi	28	1	0.0000000000	0.0000000000	Kalamitsi, Ionian Sea, Greece	1991/4/0 - 1991/4/0	On map		

Number of datasets: 141

Fertig

Select Area: Full map

Select Theme: Satellite Images

▣ SST 08/2004
□ Parity 08/2004
□ Chlorophyll 08/2004
▣ Stations
□ Station labels
□ Places

197 [Stations]

40.85, 32.54
Zoom 4111 km

Based on ALOV Map (C) 2003

Fig. 2. MedOBIS system interface: selection of survey stations using search capability based on species' scientific names (information on >3000 species assembled for the eastern Mediterranean and Black Sea), superimposed on a map of sea surface temperatures. The figure quality corresponds to that of the computer screen image

Finally, the dissemination phase is based on ALOV Map (www.alov.org/), a corporate project of ALOV Software and the Archaeological Computing Laboratory of the University of Sydney, Australia (Johnson 2004). ALOV Map is a free portable Java application for the publication of vector and raster maps on the Internet and for interactive viewing on web browsers. Its appearance and features can be controlled by XML files, by JavaScript, or by custom Java classes. Highly customised web pages can be created with this tool, including links from map objects to websites, querying attribute data and many interaction features for the user such as navigation, time filtering and working with multiple layers and thematic maps.

ALOV Map supports various data formats as sources, amongst them raster images, shapefiles and SQL (Structured Query Language) databases. A dedicated interface also allows the conversion of shapefiles to a database-readable format. The spatial data, which can thus be stored along with the biological data and taxa, can be referenced to their geographic location by performing SQL queries.

MedOBIS is based on a server running a suite of server application software programs, which are almost entirely open source. One of the benefits deriving from this type of software is that the data will not be locked up within a proprietary system. Consequently, the development of the software will not be governed by a single developer, and the ability to enhance the software and to provide interfaces for other similar projects focussing on the exchange of data will be unhindered. Apache (www.apache.org/), PHP (www.php.net), MySQL (htttp://www.mysql.org) and Tomcat (http://jakarta.apache.org/tomcat/) are components of the architecture. While ALOV Map itself is not open source, there has been a declaration from the producer that this is due to take place in the near future.

MedOBIS makes use of ALOV Map's client server mode. In this case, the connection between the applet and the database is managed by a servlet container hosted by Tomcat. This architecture allows an incremental loading of data to the client side, reducing download time and network traffic. PHP scripts are invoked to query the database; these search the data, display the results in a separate window and invoke a JavaScript code, which marks the matching stations on the map (Fig. 2).

Three approaches to obtaining information have been implemented to satisfy the need for customised queries. (1) For spatial queries, the user may select areas or single stations on the map and choose a predefined theme to obtain more information about the biological and environmental information available for these areas. (2) Taxonomic queries allow a search according to species names; the search mode displays a distribution pattern on the map and a detailed results window, which offers the user further navigation, metadata information and downloading possibilities, thus allowing the incorporation of additional environmental data. (3) Finally, an advanced interface is used to specify the request and even the output by taking into account environmental parameters, higher taxonomic groups, or certain time periods.

Additionally, a mailing list has been established that runs in parallel with the web site. It delivers information on marine biodiversity issues to approximately 600 Email addresses distributed not only in the eastern Mediterranean and Black Sea regions, but in other European and overseas countries as well. The mailing list (medobis@hcmr.gr) is also concepted to act as an electronic forum for the scientific community of the region and to contribute to the exchange of ideas and to the formulation of new projects in the future. The initiative is currently supported by the National Excellence Project of the Hellenic Centre for Marine Research on Marine Biodiversity. The centre shoulders responsibility for the sustainable maintenance of the MedOBIS initiative in the future, as part of its data management policy. This fact in itself ensures the sustainability of the MedOBIS system, which will be further developed through the introduction of online tools, metadata management, data entry user interfaces and customised data output.

Acknowledgements. This project was funded by the Niarchos Foundation through a grant to the 'Census of Marine Life' as part of its Ocean Biogeographic Information System and by the National Excellence Project of the Hellenic Centre for Marine Research in Marine Biodiversity. The authors wish to thank Mrs. M. Eleftheriou for critical reading of the manuscript.

LITERATURE CITED

Arvanitidis C, Bellan G, Drakopoulos P, Valavanis V, Dounas C, Koukouras A, Eleftheriou A (2002) Seascape biodiversity patterns along the Mediterranean and the Black Sea: lessons from the biogeography of benthic polychaetes. Mar Ecol Prog Ser 244:139–152

Babu MN (2003) Implementing Internet GIS with Java based Client/Server Environment. In: Proceedings of Map Asia Conference, 13–15 October 2003, Kuala Lumpur. (available at: www.gisdevelopment.net/technology/gis/mq03230.htm)

Costello MJ (2000) Developing species information systems: the European register of marine species. Oceanography 13:48–55

Costello MJ, Emblow C, White R (eds) (2001) European register of marine species. A check-list of marine species in Europe and a bibliography of guides to their identification. Patrimoines Naturels, Vol 50, Publications Scientifiques, Muséum national d'histoire naturelle, Paris, p 1–463

Costello MJ, Emblow C, Bouchet P, Legakis A (2006) Gaps in knowledge of marine biodiversity and taxonomic resources in Europe. Mar Ecol Prog Ser 316:257–268

Decker CJ, O'Dor R (2002) A census of marine life: Unknowable or just unknown? Oceanol Acta 25:179–186

ESRI (Environmental Systems Research Institute) (1994) ARC macro language. ESRI Press, Redlands, CA

Johnson I (2004) Putting time on the map: using TimeMap for map animation and web delivery. GeoInformatics 7:26–29

Laitinen S, Neuvonen A (2001) BALTICSEAWEB: an information system about the Baltic Sea environment. Adv Environ Res 5:377–383

Marshall J (2002) Developing internet-based GIS applications. GIS India 11:16–19

Schalk PH (1998) Management of marine natural resources through by biodiversity informatics. Mar Policy 22:269–280

Tsontos VM, Kiefer DA (2002) The Gulf of Maine biogeographical information system project: developing a spatial data management framework in support of OBIS. Oceanol Acta 25:199–206

Valavanis VD (2002) Geographic information systems in oceanography and fisheries. Taylor & Francis, London

Valavanis VD, Georgakarakos S, Haralabous J (1998) A methodology for GIS interfacing of marine data. In: Proceedings of GIS PLANET 98: international conference and exhibition on geographic information, Lisbon. Imersiva CD-ROM, www.imersiva.com

Valavanis VD, Georgakarakos S, Kapantagakis A, Palialexis A, Katara I (2004a) A GIS environmental modelling approach to essential fish habitat designation. Ecol Model 178:417–427

Valavanis VD, Kapantagakis A, Katara I, Palialexis A (2004b) Critical regions: a GIS-based model of marine productivity hotspots. Aquat Sci 66:139–148

Valavanis VD, Katara I, Palialexis A (2005) Identification of mesoscale thermal fronts using satellite imagery and GIS. Int J Geogr Inf Sci 19:1131–1147

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway,

Submitted: February 14, 2005; Accepted: September 6, 2005
Proofs received from author(s): June 14, 2006

Vol. 316: 231–238, 2006

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

Integrating information on marine species identification for fishery purposes

Jordi Lleonart*, Marc Taconet, Michel Lamboeuf

FAO Fisheries Department, Viale delle Terme di Caracalla, 00100 Rome, Italy

ABSTRACT: Species identification for fishery purposes has been the subject of a major Food and Agriculture Organization (FAO) program since the 1960s. One of the main objectives is to improve catch statistics through accurate species identification. A number of guides (geographical), catalogues (taxonomic) and species synopses have been produced as hard copy, and, more recently, most of these publications have become freely available on the Internet. Species fact sheets are a new electronic product with a database structure integrated in the Fisheries Global Information System (FIGIS). FIGIS interconnects species information with many other types of information related to fisheries (statistics, stocks inventories and assessment reports, fisheries inventories, fishing techniques, fisheries management systems, introduced species, cultured species etc.) and with a wide range of services available from other FAO systems (virtual document library, mapping library, legislation library, scientific abstracts). FIGIS achieves these features thanks to a flexible 3-tier architecture based on open-source software (Java, XML, XSL, HTML), a metadata framework based on international standards, formal institutional partnerships for information sharing, and the exploitation and display of web services. Several gaps in geographical and taxonomical coverage have been determined; these are mainly located in South America and concern several taxa (particularly crustaceans) and some fish families of paramount importance to fisheries. Other species identification tools to address multispecies and ecosystem modeling are also needed. Finally, optimization of the worldwide community efforts in generating and sharing taxonomically related knowledge in a global network is a current challenge calling for an urgent solution.

KEY WORDS: FAO · Species identification · Databases

INTRODUCTION

The Food and Agriculture Organization (FAO) of the United Nations (UN) leads international efforts in defeating hunger. Serving both developed and developing countries, the FAO is also a major source of knowledge and information. It helps developing countries and countries in transition to modernize and improve agriculture, forestry and fishery practices and to ensure good nutrition for all. The mission of the Fisheries Department of the FAO is to facilitate and secure the long-term sustainable development and utilization of the world's fisheries and aquaculture resources.

Species identification is a major fisheries issue, which has been recognized since the 1960s. The preamble of the very first species identification guide (Fischer 1973) states, 'It is hoped that the use of this new work tool will contribute to the improvement of national and regional fishery statistics and will facilitate fishery resources survey work, sampling schemes and fishery activities in general.'

The Species Identification and Data Program (SIDP) was initiated in the early 1970s to improve the quality of fisheries data collection by species through reliable species identifications in the field, in particular in developing areas and countries. Nowadays, this objective also serves the requirements of the code of conduct for responsible fisheries (FAO 1995), which provides a new framework to integrate information on biodiversity, species introductions and protection of endangered species. Hence, not only commercial spe-

*Email: jordi.lleonart@fao.org

cies are the subject of species identification publications, but also species with indirect relevance to fisheries, such as marine mammals or turtles.

When the 1995 code of conduct for responsible fisheries was approved as a basis for policies aimed at sustainable fisheries, a major need for reliable, high-quality and relevant information on world fisheries was identified. In 1999, the Fisheries Department launched the development of a global network of integrated fisheries information (FIGIS—Fisheries Global Information System; www.fao.org/fi/figis) to address this need.

FIGIS is an information management tool designed (1) to promote policy change towards the sustainable use of the world's fishery resources by highlighting major issues, presenting possible solutions and providing the best scientific information available, (2) to offer a single and unique entry point to strategic data, information, analyses and reviews of issues and trends on a broad range of fisheries subjects and (3) to provide quality-controlled, harmonized, streamlined and comprehensive information. FIGIS, as an information management tool, interconnects a network of subsystems, some of which are made accessible through partnership arrangements with other institutions.

NECESSITY TO IMPROVE SPECIES IDENTIFICATION AND FISHERY STATISTICS

Correct identification to the species level is necessary for most biological studies; however, higher taxonomic levels have also been used in ecological and fisheries analyses (Myers & Worm 2003, Pauly et al. 2003). Taxonomy is facing a prestige crisis (Godfray 2002), and the obvious connection between the increasing concern about declining biodiversity and the need for species identification is not always recognized (Boero 2001).

One of the key problems that makes fisheries management difficult is the erroneous and/or imprecise identification of the exploited species. This leads to an ill-defined or inappropriate species catch attribution, evidenced by the significant quantities of catches that are reported only in higher taxonomic level groupings in the statistics. Every year, countries are requested to provide the FAO with their catch statistics by species and fishing areas. The FISHSTAT Plus Capture production database contains catch data since 1950 (FAO 2004). It includes 1347 'species items' (actual species or groups of species). It should be noted that the number of species items has grown continuously, i.e. in 1990 the number of items was 995 (L. Garibaldi pers. comm.). For the purpose of this study the items have been divided into 4 levels of taxonomic groups: (1) species, e.g. *Sardina pilchardus*; (2) genus, e.g. *Dentex* spp.; (3) family, e.g. Congridae; and (4) higher taxonomic groups (HTG), e.g. Gadiformes, Bivalvia, Osteichthyes.

Table 1. Taxonomic aggregation of average world catch data (in metric tons) from 1950 to 2002. HTG: higher taxonomic groups; SI: species items

	SI	Tons	Percent of catch	Tons SI^{-1}
Species	1008	42 785 479	66.45	42 446
Genus	180	5 485 170	8.52	30 473
Family	109	2 554 751	3.97	23 438
HTG	50	13 557 787	21.06	271 156
Total	47797	96 292 408		71 487

In 2002, more than 60 % of the catch was attributed to 1008 single, identified species, but more than 20 % of the catch was included in the too general HTG category (Table 1), attributed to only 50 items. Caddy & Garibaldi (2000) found that 65.9 % of the total capture production reported to the FAO for 1996 was at the species level, but also observed a great difference between temperate areas, with 90 % at the species level, and tropical areas, where it was often lower than 40 %.

A study of the historical series of world capture fishery production statistics shows that species resolution is declining. The number of catches reported at the species level is decreasing, while the trend for reports of aggregated groups is increasing (Fig. 1); these trends are statistically significant in all cases (Table 2). An important source of imprecision lies in the statistics of some SE Asian countries. For example, data from China, Vietnam, Myanmar, Indonesia and eastern Thailand account for 62 % of all the reported catch attributed to the HTG. After removing data from these countries, the % of HTG dropped, and the profile became even, with a nearly flat trend. With regard to species, the negative slope is attenuated, but still remains significant.

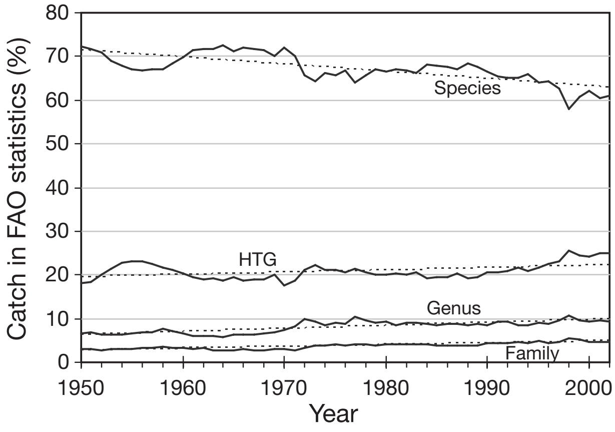


Fig. 1. Levels of catch identification in FAO statistics from 1950 to 2002 (solid lines: observed values; dotted lines: linear regressions; HTG: higher taxonomic groups)

Table 2. Linear correlations between group allocation of catches for both annual and total catch. All values are significant (values higher than 0.27 are significant at α = 0.05 and df = 51). The linear correlation between annual and total catch is 0.9855

	Annual	Total catch
Species	–0.7442	–0.6518
Genus	0.8053	0.7738
Family	0.8933	0.8264
HTG	0.4197	0.3004

The decrease in capture fishery statistics species identification (with annual and/or total catch) is worrying. This can be attributed to an increase in catches grouped together as 'Osteichthyes, marine fishes nei' (nei = not elsewhere included), the number of which are constantly increasing, and to groups such as 'Crustacea' and 'Mollusca' that have increased in importance since 1990. However, some positive trends should be pointed out. For example, catches of 'sharks, rays and chimaeras' have been stable since 1996, at about 0.8 million metric tons, but the number of items reported at species level has increased from 45 to 95, indicating a remarkable improvement in species identification (FAO 2005). More generally, the continuous increase in the number of species items is also a positive indicator. Some regions have shown marked progress in species identification in the statistics, as in the eastern Central Atlantic, where the percentage of catches reported at species level increased from 43% in 1970 to 65% in 2002, while HTG identifications dropped from 46 to 28% during the same period (CECAF 2004).

FAO AND SPECIES IDENTIFICATION

The objectives of the SIDP are (1) to improve the identification of marine organisms of actual and potential interest to fisheries; (2) to provide and disseminate tools to facilitate species identification in fisheries and, in so doing, improve fisheries data quality and (3) to provide a global and coherent system of scientific and common nomenclature. Priority is assigned to resources of major commercial importance that are considered threatened and to developing regions facing difficulties in species identification. SIDP relies on a network of >100 highly reputed taxonomists and scientists, each a specialist in his/her group, to process and validate the species identification information to be published.

SIDP produces 5 species identification collections, a large number of which are available at www.fao.org/fi/sidp.

Regional guides. Comprehensive, coded, annotated and illustrated inventories of the species in some regions of the world are available. These include dichotomous identification keys and are based on contributions of a large group of taxonomists and fishery technicians. They are available as paper publications, and some are also available at the above website. So far, 8 such guides have been published (Table 3).

Field guides. Guides of commercial species containing fish landings for individual countries or groups of countries are also available. They are illustration-based, with pictorial keys to families, a minimum of text and include common names. They are particularly aimed at national data collectors in need of quick identification of species in markets and landing places for the specific purpose of improving statistical and other fisheries data by species. So far 17 field guides have been published (Table 4). They are available as paper publications, and a few are also available at the SIDP website.

Catalogues. Worldwide inventories of taxonomic groups have been published. The scope of a catalogue is to deal with 1 or several taxonomic groups (class, order, family or subfamily) and to include all known species of the group around the world. They contain dichotomous identification keys: drawings, detailed descriptions, synonyms and world distribution. So far, 22 catalogues have been published. They are available both as paper publications and at the SIDP website The catalogues cover, partially or completely, 5 main

Table 3. List of regional guides

Editor(s)	Year	Region covered (publication language)
W. Fischer & P. J. P. Whitehead	1974	Eastern Indian Ocean and western Central Pacific (English)
W. Fischer, G. Bianchi & W. B. Scott	1981	Eastern Central Atlantic (English & French)
W. Fischer & G. Bianchi	1984	Western Indian Ocean (English)
W. Fischer & J. C. Hureau	1985	Southern Ocean (English, French & Spanish)
W. Fischer, M.-L. Bauchot & M. Schneider	1987	Mediterranean and Black Sea (French)
W. Fischer, F. Krupp, W. Schneider, C. Sommer, K. E. Carpenter & V. H. Niem	1995	Eastern Central Pacific (Spanish)
K. E. Carpenter & V. H. Niem	1998	Western Central Pacific (English)
K. E. Carpenter	2002	Western Central Atlantic (English)

Table 4. List of field guides

Author(s)	Year	Country covered (publication language)
M.-L. Bauchot & G. Bianchi	1984	Madagascar (French)
G. Bianchi	1985	Pakistan (English)
G. Bianchi	1985	Tanzania (English)
G. Bianchi	1986	Angola (Portuguese)
M. Bellemans, A. Sagna, W. Fischer & N. Scialabba	1988	Senegal and Gambia (French)
W. Fischer, I. Sousa, C. Silva, A. De Freitas, J.-M. Poutiers, W. Schneider, T. C. Borges, J. P. Féral & A. Massinga	1990	Moçambique (Portuguese)
W. Schneider	1990	Gulf of Guinea (English & French)
D. H. Eccles	1992	Tanzania [freshwater] (English)
F. Cervigón, R. Cipriani, W. Fischer, L. Garibaldi, M. Hendrickx, A. J. Lemus, R. Márquez, J.-M. Poutiers, G. Robaina & B. Rodriquez	1993	Northern coast of South American (English & Spanish)
G. H. P. De Bruin, B. C. Russell & A. Bogusch	1995	Sri Lanka (English)
W. J. Rainboth	1996	Cambodian Mekong (English)
C. Sommer, W. Schneider & J. M. Poutiers	1996	Somalia (English)
K. E. Carpenter, F. Krupp, D. A. Jones & U. Zajonz	1997	Kuwait, eastern Saudi Arabia, Bahrain, Qatar and the United Arab Emirates (English)
D. Lloris & J. Rucabado	1998	Morocco (French)
G. Bianchi, K. E. Carpenter, J.-P. Roux, F. J. Molloy, D. Boyer & H. J. Boyer	1999	Namibia (English)
R. Bonfil & M. Abdallah	2004	Sharks and rays of the Red Sea and Gulf of Aden (English)
F. Serena	2005	Sharks and rays of the Mediterranean (English)

groups: fish, crustaceans, cephalopods, turtles and marine mammals. The catalogues published to date cover: shrimps and prawns, marine lobsters, cephalopods, marine turtles, marine mammals, sharks, and bony fish (Scombridae, Istiophoridae, Xiphiidae, Lutjanidae, Clupeoidei [Chirocentridae, Clupeidae, Pristigasteridae and Engraulidae], Caesionidae, Lethrinidae, Gadiformes, Nemipteridae, Sillaginidae, Gempylidae, Trichiuridae, Epinephelinae, Glaucosomatidae, Ophidiformes, Merlucciidae and Polynemidae).

Synopses. Synopses are comprehensive reviews of current knowledge on species of aquatic organisms of present or potential economic interest. Every synopsis covers all aspects of a single species: taxonomy, distribution, biology, fishery, utilization etc. They are available as paper publications, and a few are also available in electronic form.

Fact sheets on aquatic species. These are only available in electronic form through the SIDP website. The priorities for publishing species fact sheets are based on several criteria: species with a high level of world catch or a high revenue level, and those of importance with regard to biodiversity and conservation. At the end of 2004, 547 species fact sheets were available in FIGIS: 111 on sharks, 235 on bony fish, 42 on crustaceans, 9 on cephalopods, 24 on shells, 117 on mammals, 3 on turtles, 4 on algae, 1 on corals and 1 on sea urchins. Their dynamic generation and organization is further described in more detail.

NEW INTERNET-BASED TECHNOLOGY FOR DISSEMINATION AND EXCHANGE OF SPECIES IDENTIFICATION

It is not difficult nowadays to find more than 50 different websites on the Internet dealing with species identification on many levels, i.e. taxonomic (all groups, fish, cephalopods, crustaceans etc.), geographical area (worldwide or regional), special characteristics (e.g. invasive, endangered), use (e.g. aquariology, fisheries, aquaculture, SCUBA diving, food safety), parts of animals (e.g. otoliths), or life history (e.g. larvae). This reflects a great heterogeneity of data sources and products. However, only some of them can currently be recognized as references.

FishBase, a tremendously content-rich encyclopedia, is a much used source, although limited to fish in a taxonomic sense. The Integrated Taxonomic Information System (ITIS) appears to be the emerging web-based reference for standard taxonomic information on marine life (and more). Having online access to such a central register based on a commonly used taxonomic system across marine life, with an ever-increasing registry of living organisms managed under the Global Biodiversity Information Facility (GBIF), is essential to underpin biological databases and to enable web-based interoperability between information systems.

The Ocean Biogeographic Information System (OBIS) provides a good example of what a simple web-based system protocol can achieve in dynamically mapping species occurrence worldwide. This could be a tool of great interest for cross-checking and validating species distribution maps and, where data allow, generating some spatial index of probability of occurrence.

THE FIGIS SYSTEM: AN ANSWER TO GROWING INFORMATION REQUIREMENTS

Objectives. The role of FIGIS as an information tool is to support the FAO framework in its implementation of the Code of Conduct for Responsible Fisheries (CCRF; FAO 1995). The approach is fisheries-centered, with a focus on exploitation, usage and management of fishery resources. By offering a single entry point to comprehensive, reliable, high-quality and relevant information on the state of world fisheries, FIGIS should contribute to the promotion of policy changes towards the sustainable use of the world's fishery resources.

The various FAO fisheries programs (information and statistics, policy, resources, utilization and technology) have historically developed their own information bases to support their thematic analyses. FIGIS's primary objective was to consistently interconnect these databases. As a result, FIGIS currently handles 30 databases spread across 20 distinct information domains, such as statistics, aquatic species (SIDP fact sheets), introduced species, stocks, fisheries, management systems, institutions, laws, fishery country profiles, cultured species, glossaries, bibliographic references, news and events, etc., as well as standard classifications stored as reference data.

Content. The content of FIGIS should be understood as an information pyramid with various layers targeting distinct audiences. The top layer, aimed at policy makers, is presented in the form of policy notes, highlighting major fishery issues, presenting perspectives and possible solutions. Consistent with best practices as enshrined in international agreements, these policy notes and strategic summaries are systematically supported by the best scientific information available, thanks to links established to more technical information located in the lower layers of the information pyramid. Hence, at the base of the pyramid, FIGIS integrates detailed technical information of interest to analysts or scientists, such as fishery statistics, species identification, introduced species, fishing technology, description of geographical areas, of fishery management institutions, etc. In the middle of the pyramid, it facilitates the development of new information products, such as stocks or fisheries inventories, or dynamic country profiles. Responding to the best practices, these products enable users to trace source documents from global syntheses or reviews such as the Review of the state of world marine fishery resources (FAO 2005). Beyond its ability to integrate existing information bases, this middle layer constitutes the major novelty of FIGIS, as highlighted in the next section.

NEW OPPORTUNITIES OFFERED BY FIGIS

New databases for improved management of fishery resources. The newly developed marine resources and fisheries inventories databases well illustrate the role of FIGIS and the issues to which it responds: fishery statistics are generally highly aggregated and excessively based on catch or trade, resulting in minor fisheries being overlooked in policy making, despite the fact that collectively they can play a considerable role in the national economy and human livelihoods. An inventory of resources and fisheries would help to estimate the contribution of these poorly monitored fisheries and encourage policy-making to take into account the communities that depend upon these fisheries. The fisheries inventory also provides the backbone for linking other information in FIGIS to characterize fisheries management and its effectiveness: governance systems, monitoring indicators, management measures, scientific advice and related management actions, and the response of fisheries resources. As of December 2004, this inventory includes about half of the global coverage, with 1900 stocks and 2300 fisheries enumerated. The current status can be found at www.fao.org/figis/servlet/static?xml=STF_proj.xml&dom=org&xp_nav=4,2.

Innovative information products conveying evolving knowledge. The FIGIS fact sheet, a fundamental FIGIS product, is a textual synthesis of information supported by tables, graphics and maps. Reached through the gateway pages, fact sheets are designed to present various characteristics within each broad fishery subject. Fact sheets implement the aforementioned system capacity; thanks to control exerted by a metadata level, they contain dynamically assembled information from various databases and present them in a homogenized and structured layout. Thus, in a controlled manner, fact sheets elaborate new knowledge while assembling information from previously disconnected and dispersed sources.

A species fact sheet includes the following features. A heading section presents the institutional data ownership together with the procedures, sources and methods used to compile the information. The cover page, together with the species scientific name, provides the title elements for web-page citations. The identification section consists of the species scientific name, together with other identification elements, such as image(s) of the species, and different official names and standard codes. The main section deals with a number of standard topics, including diagnostic features, habitat and biology, size, geographic distribution, interest to fisheries and conservation status. Finally, the sources and bibliographic section contains bibliographic references used/consulted to draw up the different parts of the fact sheet.

The core textual elements of the species fact sheet are provided by the SIDP program, often from information published in FAO catalogues, or specifically compiled and written by specialists, since the catalogues do not cover all important commercial groups.

One of the key features of fact sheets is the dynamic link to other databases; data, graphs, maps and links are seamlessly and dynamically assembled from other sources thanks to the different levels of interoperability possible with the FIGIS system. In the identification section, all the identification data elements (except images) are extracted on the fly from the Aquatic Sciences Fisheries Information System (ASFIS) database hosted in FIGIS. Thanks to an agreement with FishBase on interfacing databases using standard species codes, each paragraph of the fish species fact sheet links directly to the relevant data in FishBase and presents these data in a separate window. The same process applies to the links to CephBase for cephalopods and to the Information System for the Promotion of Aquaculture in the Mediterranean (SIPAM) for aquatic animals cultivated in the Mediterranean region. The species distribution map is dynamically generated from the GIS component of FIGIS. The interest to fisheries topic contains 1 or 2 statistical graphs (depending on whether the species is captured and/or cultivated) dynamically generated from the FAO global capture fishery and aquaculture statistics online database managed by FIGIS. It also links to dynamically generated lists of records from different domains (fishing techniques, fish stocks, fisheries, introduced species) indexed with this species. The bibliography section dynamically triggers requests to the FAO virtual library when reference is made to FAO publications. Similar dynamic queries will be directed in the near future to the Aquatic Sciences and Fisheries Abstracts (ASFA) database.

FIGIS-ENABLING TOOLS AND MECHANISMS

FIGIS technology. Designed as a 3-tier architecture web-based system, FIGIS is based on a relational database management system and open-source programming languages (Java, XML, XSL and HTML). As such, it is proprietary-free software and a platform-independent application. It also integrates a GIS component, allowing systematic data geo-referencing. FIGIS is open to interoperability with external systems thanks to the adoption of international metadata standards, including Dublin Core, Agricultural Metadata Element Set (AGMES), Resource Description Framework Schema (RDFS), Ontology Web Language (OWL) and classifications of the International Standards Organization (ISO), the United Nations, the Coordinating Working Party on fishery statistics (CWP). This technology facilitates a flexible, scalable and distributed architecture, providing the ability to dynamically assemble data from distributed sources.

The metadata framework. Modules for marine resources, fisheries and other purposes (fishing techniques, management systems, etc.) are all new and complex information concepts. Handling these concepts consistently at the global level raises the challenge of coherent definitions, which is one of the purposes fulfilled by FIGIS metadata. Furthermore, FIGIS metadata provide a model for these complex concepts, based on elementary building blocks at the bottom of the information pyramid, such as aquatic species, gear type, vessel type, water area, or country, all of which use standard international classifications. Geographic entities included in the metadata are backed by a GIS component, thus enforcing systematic data georeferencing against standard, widely distributed GIS shape files. In utilizing this metadata standard framework, FIGIS ensures consistent and accurate links between the databases that are integrated to increase the value of the disseminated information. It also enables the dissemination of maps showing the location of stocks or fishery units, species distribution, or the spatial distribution of catch statistics.

Protocols for data sharing. Protocols implemented by FIGIS range from providing focused gateways to external systems such as FishBase, through dynamic and seamless assembly within FIGIS information pages located in partner systems, to the provision of content management system services. The latter include an upload service for registered users or systems willing to load their information in the FIGIS database from a range of formats (Excel, CSV, or XML files for statistical data, XML for metadata or textual information) and an online editing service for more simple direct inputs through HTML forms. Thanks to the metadata framework, data entry mechanisms exert tight control checks, verifying correct indexing over reference data. Here, the dream of information flow streamlining begins to become a reality.

Partnership arrangements. A key principle in developing FIGIS is to ensure that information is sustainable, quality-controlled, updated and provided by the most authoritative source. This is achieved through the development of partnerships which enable the sharing of information within a global network. The aforementioned global inventories of marine resources and fisheries provide the backbone of a Fishery Resources Monitoring System (FIRMS) currently being developed in close collaboration with regional fisheries organizations. FIRMS is a formal partnership arrangement adopted in February 2004 aiming at the systematic assembly of quality-controlled information on status and trends of fishery resources. It currently involves 8 regional fishery organizations willing to contribute information to the system according to their own mandates.

ISSUES RAISED BY THE GLOBAL INTERNET UNDERTAKING

Diversity of definitions/views worldwide. The challenge of harmonizing stocks or fisheries status reports worldwide lies in the ability to propose acceptable definitions and topic trees to all participating institutions. When one asks 'What is a fishery?', the global community gives numerous responses and definitions. This variability is principally a question of scale (which aggregation level is being considered?), coupled with the disciplinary angle used when defining a fishery. In acknowledgement of the most frequent approaches to fisheries analysis and reporting, the retained definition emphasizes that the fisheries concept stresses the usage and management perspective of fishery resources. According to the FIGIS definition, extending that of the FAO Fisheries Glossary (http://fao.org/fi/glossary), 'a fishery is an activity leading to harvesting fish within the boundaries of a defined area; the fishery concept fundamentally gathers indication of human fishing activity including from the management, biological/environmental and technological view points'.

Towards more complex requirements. Integrating more databases means facing additional requirements. The records in the introduced species database not only include references to standard species, but also to subspecies, strains, or hybrids. In contrast, many fisheries databases refer to species groups, or to species using common local names. The emerging demand for traceability of fishery products against their environment of origin implies the ability to develop and manage systematic lists of geo-referenced environments. Enabling consistent cross-referencing on species or geographic components on a global level requires major effort and close coordination.

An immense potential to handle carefully. Assembling information from heterogeneous databases presents immense processing and modeling challenges. Original sources are compiled in different contexts, following distinct methodologies and referring to different terminologies or definitions. Enabling searches on integrated sources for selective extraction of information or applying algorithms generating summary indicators may produce misleading knowledge by presenting parts of the original information out of context. This is where frameworks on information sharing policy, agreed upon by stakeholders, will play an increasingly important role in the near future.

DISCUSSION AND CONCLUSIONS

Species identification work in the FAO has essentially been aimed at improving fisheries statistics. Gaps in publication for many regions or taxonomic groups still remain. (1) No guides exist for FAO Fishing Areas 41 and 87 (South America), which correspond to important fishing zones and developing countries for which no other regional guides are available. (2) Very important taxonomic groups, from a fisheries point of view, are not yet covered by FAO catalogues, e.g. batoids, flatfishes, sparids and carangids. (3) A database on crustaceans is needed, though some information on crustaceans is already available in FIGIS and other sources. (4) There is a need for information on non-commercial taxonomic entities that may have an impact on, or be impacted by, fisheries (jellyfish, corals, etc.).

Fisheries science requires taxonomic stability for the well-known species and quick taxonomic updating for new target species or species affecting fisheries in some way. Newly exploited species, such as those exploited on seamounts, or species groups, such as sharks, for which the conservation status changes to endangered require guides for accurate short-notice identification.

The requirements for species identification have changed. The greater challenge posed by the poor state of world fishery resources brings with it an ever broadening need for information. The multifaceted fisheries approach of the FAO constitutes a response whereby aquatic species information occupies a crucial position: species are targets of fishing techniques; species define key dynamic population features of fishery resources and stocks; and species determine the structure of biogeographic components of multispecies fisheries. High-quality quantitative and qualitative reporting requires accurate taxonomic identification and the availability of a thesaurus of vernacular names in multiple languages for species and species groups, as well as ways to connect scientific names to vernacular names.

Progressive implementation of the ecosystem approach to fisheries management leads to emerging needs that can be categorized into 3 groups: (1) analysis of trophic relationships, which requires the ability to identify species from parts of organisms, such as otoliths, scales, or parts of crustacean carapaces, etc.; (2) libraries for ecological modeling parameters; and (3) synthetic products from ecological modeling, such as biodiversity indicators in multispecies assemblages, faunistic changes, or identification of ecosystem units. Faunistic changes, such as anthropogenic species introductions, or variations in the distribution area due to climatic changes, should be rapidly recorded in the databases. The Mediterranean and Black Sea, with the effects of Lessepsian species, tropicalization (Quignard & Tomasini 2000) and water-ballast introductions with the dramatic consequences they have for fisheries, such as the ctenophore *Mnemiopsis leidyi* (Zaitsev & Öztürk 2001), is an illustration of this need. The CIESM atlas of exotic species is a significant contribution in this regard (Galil et al. 2002, Golani et al. 2002, Zenetos et al. 2003).

The needs identified are immense, and the Internet opens new but challenging horizons. Internet technologies are certainly a way to 'reinvent' taxonomy (Gewin 2002, Godfray 2002), and one question could be 'Are all of these databases redundant?' or, more importantly, 'Are there any taxonomic groups absent from the Internet?' The main issue is certainly how to optimize world community efforts to generate and share taxonomically related knowledge in a global network and, consequently, how to construct globally accessible knowledge. The FIGIS experience of a web-based system integrating knowledge from distributed web resources shows that the key to success is to adopt international metadata standards and forge agreements on authoritative lists, classifications and coding systems. The greatest challenge related to this approach is the ability of the various interconnected databases to share compatible semantics. This is further complicated by the fact that the adoption of metadata standards is usually discussed within communities of similar interest, and the need to interrelate metadata standards among dissimilar communities sharing a few common dimensions remains a major challenge. Ongoing research efforts are being coordinated by the FAO to demonstrate how ontological methodologies may provide responses to this issue and enhance semantic interoperability in fishery information systems (Gangemi et al. 2004). Globally, the role of each party has to be clearly defined according to the party's mandate. FIGIS, supporting FAO's efforts to promote sustainable and responsible use of fishery resources, offers the biogeographic community Internet services that are directly related to the FAO's institutional role. Web services are available to clients for dynamic retrieval of international classifications and of reference lists maintained by the FAO, or of XML-based products, such as species fact sheets or fishery statistics.

Acknowledgements. The authors are very grateful to Mr. L. Garibaldi for his comments and advice on catch statistics, and to Mr. F. Carocci for providing the cartography.

LITERATURE CITED

Boero F (2001) Light after dark: the partnership for enhancing expertise in taxonomy. Trends Ecol Evol 16(5):266

Caddy J, Garibaldi L (2000) Apparent changes in the trophic composition of world marine harvest: the perspective from FAO capture database. Ocean Coast Manag 43:615–655

CECAF (Committee of the Eastern Central Atlantic Fisheries) (2004) Regional implementation of the strategy for improving information on status and trends of capture fisheries. In: Report of the 17th session of the Fishery Committee for the Eastern Central Atlantic. Dakar, Senegal, 24–27 May 2004. FAO Fisheries Report No. 754, FAO, Rome

FAO (Food and Agricultural Organization) (1995) Code of conduct for responsible fisheries. FAO, Rome

FAO (Food and Agricultural Organization) (2004) FAO yearbook fishery statistics. Capture production 2002. FAO Fisheries Series No. 66, FAO Statistics Series No. 180, Vol 94/1 2002. FAO, Rome

FAO (Food and Agricultural Organization) (2005) Review of the state of world marine fishery resources. FAO Fisheries Technical Paper No. 457, FAO, Rome

Fischer W (ed) (1973) Mediterranean and Black Sea (Fishing Area 37). FAO species identification sheets for fishery purposes. FAO, Rome

Galil B, Froglia C, Noël P (2002) Crustaceans: decapods and stomatopods. In: Briand F (ed) CIESM atlas of exotic species. 2. CIESM Publishers, Monaco

Gangemi A, Fisseha F, Keizer J, Pettman I, Taconet M (2004) A core ontology of fishery and its use in the fishery ontology service project. In: Gangemi A, Borgo S (eds) First International workshop on core ontologies. EKAW Conference CEUR-WS, Vol 118. Available at http://sunsite.informatik.rwth-aachen.de/Publications/

Gewin V (2002) All living things, online. Nature 418:362–363

Godfray HCJ (2002) Challenges for taxonomy. The discipline will have to reinvent itself if it is to survive and flourish. Nature 417:17–19

Golani D, Orsi-Relini L, Massutí E, Quignard JP (2002) Fishes. In: Briand F (ed) CIESM atlas of exotic species. 1. CIESM Publishers, Monaco

Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. Nature 423:280–283

Pauly D, Alder J, Bennett E, Christensen V, Tyedmers P, Watson R (2003) The future for fisheries. Science 302: 1359–1361

Quignard JP, Tomasini JA (2000) Mediterranean fish biodiversity. Biol Mar Mediterr 7(3):1–66

Zaitsev Y, Öztürk B (eds) (2001) Exotic species in the Aegean, Marmara, Black, Azov and Caspian Seas. Turkish Marine Research Foundation, Istanbul

Zenetos A, Gofas S, Russo G, Templado J (2003) Molluscs. In: Briand F (ed) CIESM atlas of exotic species. 3. CIESM Publishers, Monaco

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: January 14, 2005; Accepted: December 15, 2005
Proofs received from author(s): May 24, 2006

Vol. 316: 239–246, 2006

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles

P. N. Halpin[1,*], A. J. Read[2], B. D. Best[1], K. D. Hyrenbach[2], E. Fujioka[1], M. S. Coyne[1], L. B. Crowder[2], S. A. Freeman[2], C. Spoerri[1]

[1]Marine Geospatial Ecology Laboratory, Duke University, Durham, North Carolina 27708, USA
[2]Duke University Marine Laboratory, Beaufort, North Carolina 28516, USA

ABSTRACT: Our ability to understand, conserve, and manage the planet's marine biodiversity is fundamentally limited by the availability of relevant taxonomic, distribution, and abundance data. The Spatial Ecological Analysis of Marine Megavertebrate Animal Populations (SEAMAP) initiative is a taxon-specific geo-informatics facility of the Ocean Biogeographic Information System (OBIS) network. OBIS-SEAMAP has developed an expanding geo-database of marine mammal, seabird, and sea turtle distribution and abundance data globally. The OBIS-SEAMAP information system is intended to support research into the ecology and management of these important marine megavertebrates and augment public understanding of the ecology of marine megavertebrates by: (1) facilitating studies of impacts on threatened species, (2) testing hypotheses about biogeographic and biodiversity models, and (3) supporting modeling efforts to predict distributional changes in response to environmental change. To enhance the research and educational applications of this database, OBIS-SEAMAP provides a broad array of web-based products and services, including rich species profiles, compliant metadata, and interactive mapping services. This system takes advantage of recent technological advances in Geographic Information Systems (GIS), Internet data standards, and content management systems to stimulate a novel community-based approach to the development of a data commons for biogeographic and conservation research. To date, the global OBIS-SEAMAP database includes >1 million observation records from 163 datasets, spanning 71 yr (1935 to 2005) provided by a growing international network of data providers.

KEY WORDS: Biogeography · Marine mammals · Seabirds · Sea turtles · Oceanography · Spatial ecology · GIS · Ecoinformatics · OBIS

INTRODUCTION

Objectives

Understanding biogeographic patterns in marine systems requires integrating data from many disparate disciplines (e.g. systematics, ecology, oceanography) gathered over multiple temporal scales (e.g. seasons, years, decades) (McGowan 1990, NRC 1996, Pierott-Bults 1997). In 1997, the Alfred P. Sloan Foundation, in conjunction with the National Oceanographic Partnership Program (NOPP), initiated the ambitious 10 yr 'Census of Marine Life' (CoML) to enhance the biogeographic and ecological understanding and appreciation of marine biodiversity (Ausubel 1999). This initiative seeks to answer 3 basic questions: What used to live in the sea? What currently lives in the sea? And what will live in the sea? The CoML program includes the Scientific Committee on Oceanic Research New Technologies Working Group, which communicates awareness of advanced technologies supporting the CoML efforts and of 4 related program areas: (1) the History of Marine Animal Populations (HMAP), a synthesis of historical marine biodiversity data during the

*Email: phalpin@duke.edu

last 500 yr; (2) CoML pilot field projects, designed to test and implement novel sampling technologies; (3) the Ocean Biogeographic Information System (OBIS), an initiative to provide global, species-level, geo-referenced biogeographic data; and (4) the Future of Marine Animal Populations (FMAP) initiative, a modeling effort to determine changes in biodiversity and species distributions in response to anthropogenic impacts and climate change (Decker & O'Dor 2002). The information system role of OBIS provides a critical bridge between historic mapping, new field projects, and future modeling efforts. A recent baseline report entitled 'The Unknown Ocean' provides an assessment of the known, unknown, and unknowable in the global ocean, and summarizes the challenges and opportunities that lay ahead for the CoML program (O'Dor 2003).

A necessary first step when compiling a global baseline of marine biodiversity entails the creation of OBIS to compile, store, package, and disseminate geo-referenced biological and physical information to a broad array of users worldwide. While OBIS was envisioned as the repository of existing digital biogeographic datasets, including those originating from CoML field projects, the recent development of powerful web-based informatics and mapping tools has vastly expanded the potential research and educational applications of this initiative (Alldredge et al. 1999, Grassle & Stocks 1999).

Perhaps the most innovative aspect of the OBIS system is the planned comprehensive perspective of marine ecosystems it encompasses, by integrating information on physical properties (e.g. ocean temperature), ocean productivity patterns (e.g. chlorophyll *a* concentration), mid-trophic-level organisms (e.g. fish and squid), and top predators (e.g. large predatory fishes, marine mammals, seabirds, sea turtles). This integrative approach will enhance our understanding of which physical–biological mechanisms structure marine ecosystems, by providing simultaneous information about bottom-up (e.g. productivity) and top-down (e.g. predation) regulation of marine food webs. In addition to addressing the way entire ecosystems are structured and respond to oceanographic variability, the OBIS system will help assess the magnitude of anthropogenic impacts on marine systems. In particular, an understanding of the way marine organisms are influenced by biological and physical properties will delineate the critical habitats of protected species, and will help interpret apparent changes in population abundance by placing them in a broader oceanographic and climatic context.

The OBIS-SEAMAP program has developed an operational prototype system for the integration of oceanographic information with animal observation data. Similar systems and functionality are planned for the larger OBIS network.

Participants at CoML planning workshops repeatedly emphasized the importance of including upper-trophic marine predators in this initiative, due to their conservation status and their critical role as ecosystem-level indicators (Bradley 1999, Levi et al. 1999). In particular, the recent development of miniaturized telemetry and archival tagging technologies has facilitated the use of marine megavertebrates as autonomous sampling platforms of the marine environment, whereby researchers can integrate fine-scale behavioral information (e.g. diving) with physical environmental data (e.g. water temperature) (Stone et al. 1999). The ability to sample the 3-dimensional environment where marine megavertebrates forage at the appropriate spatial and temporal scales is providing revolutionary insights into the way these animals make a living, and is helping to delineate important migration and foraging grounds (Block et al. 2003, Welch et al. 2003).

Principal investigators at Duke University, in conjunction with a consortium of international partners, initiated the Spatial Ecological Analysis of Megavertebrate Animal Populations (OBIS-SEAMAP) initiative in 2002. The aim of this project was to assemble a global geo-referenced data repository for marine mammals, birds, and turtles, as part of the OBIS initiative. This publicly available biogeographic resource includes at-sea and colony-based absolute counts and standardized metrics of relative abundance, standardized metadata describing survey and data-processing methods, and species profiles with detailed ecological and taxonomic information.

Development of a data commons

The creation of a data commons for biogeographic and conservation research is only feasible when a sense of community exists amongst researchers, data holders, administrators, and the users of such a system. Sufficient benefits must be provided to encourage the different participants to join the community. Most researchers involved in biodiversity informatics programs recognize their value as a way to expand future scientific inquiry into questions of a spatial and temporal scope larger than any individual researcher can currently tackle in isolation. Nevertheless, the enhanced ability to seek novel scientific questions fueled by the availability of larger datasets may not suffice to stimulate some reticent data providers into contributing their data holdings. Rather, these reluctant participants may need more tangible and specific rewards. OBIS-SEAMAP has been developed with this in mind, and includes a wide array of attractive tools and services to the research and conservation communities. More specifically, to enroll data providers into the system we have devised 3 types of services:

(1) '*data management*', (2) '*value added*', and (3) '*community development*'.

Data management services are designed to facilitate the integration of the various datasets into the OBIS-SEAMAP database, by assisting with quality assurance and quality control (e.g. speed filters to identify erroneous locations along a survey track), dissemination (e.g. tools to share and disseminate research results with colleagues and founders), information technology benefits (e.g. data back-up in the OBIS-SEAMAP server), and advertising (e.g. high visibility of individual datasets and supporting citations on the public OBIS-SEAMAP website as well as the OBIS portal and other metadata clearinghouses).

Value added services enrich the datasets contributed to OBIS-SEAMAP in a variety of ways, which can include providing additional ancillary data (e.g. automatic integrated taxonomic information system, a taxonomic hierarchy for species recorded in contributed datasets), developing metadata to enhance the long-term use of the data (e.g. automated creation of mandated Federal Geographic Data Committee [FGDC]; www.fgdc.gov/clearinghouse/clearinghouse.html metadata), and merging the biogeographic information with additional environmental datasets (e.g. assigning remotely sensed environmental conditions from satellites to sightings in a given dataset).

Community development services include the OBIS-SEAMAP web tools designed to give data providers and users the ability to add content to the site (e.g. users have their own page and can post news items and announcements for public viewing), the right to manage their own datasets (e.g. a data provider can add or remove public access to a given dataset by clicking a button in a private 'mydata' page), formation of a supervising steering committee of highly respected community members, and broad exposure to the public (e.g. each dataset features links to the data provider contact pages, and includes relevant citations of published papers). In addition, users can provide access to groups of colleagues to promote collaborative efforts. To further nurture the sense of community, OBIS-SEAMAP has engaged the broader community of data providers and system users through a series of outreach activities. Namely, we have organized annual meetings with data providers and steering committee members, and have made presentations at international scientific meetings (e.g. in the fields of oceanography and of taxon-specific and geo-informatics), as well as meetings of potential user groups (e.g. marine educators, resource managers).

A critical aspect in the process of nurturing a sense of community entails addressing the concerns and needs of data providers at the onset of system design. An awareness of the apprehensions about data sharing is critical for effective system development. The best way to reconcile these disparate perspectives is for individuals to participate in the process through multiple roles. For instance, by contributing their own data to the system, the OBIS-SEAMAP developers have confronted the same proprietary (e.g. crediting funders) and scientific (e.g. ensuring faithful representation of the data) issues faced by other providers. This mutual understanding, facilitated by the trusted and personal connection between system developers and data providers, has been essential when devising approaches to protect proprietary data rights and to manage data access by the public and the scientific community.

OBIS-SEAMAP has developed terms of data use which protect the rights of data contributors without restricting the applicability of the system for a wide array of educational and conservation applications (http://seamap.env.duke.edu/about/termsofuse). All data are made available as whole datasets, with full citation and contact information from the original data providers. Complete metadata and the terms of use are bundled with the download of datasets to further promote responsible usage and proper citation.

As ecological research becomes increasingly global in scope and data intensive, biodiversity informatics programs will need to reconcile diverse national and international data access issues in a systematic fashion and in a policy arena that transcends national jurisdictions (Arzberger et al. 2004). As a data aggregator at this global scale, OBIS-SEAMAP will continue to adapt to modern 'community development' approaches and encourage good practices by others in this field.

INFORMATION SYSTEM DEVELOPMENT

The OBIS-SEAMAP system provides a wide variety of products, designed to meet the diverse needs of educators, students, resource managers, and researchers interested in marine biogeography. In particular, the web-based GIS applications make the OBIS-SEAMAP datasets widely accessible to students, researchers in less developed countries, and other users without access to expensive desktop GIS programs.

The tools used for storing, distributing, and visualizing data in OBIS-SEAMAP leverage existing software, standards, and initiatives. Specific technologies evolve rapidly, but because the framework used by OBIS-SEAMAP relies on open-standards and open-source products, the system can adapt quickly. Some of the most attractive aspects of these open-source technologies include their low cost, standards compliance, reusability, and customizable nature. The use of open standards, such as the Open GIS Consortium (OGC) standards, is especially important to promote the usage and interoperability of marine animal observation data between different software platforms and web service applications.

Overall strategy and thematic focus

The scope of this project has required engaging a wide array of stake-holders, including owners of existing datasets, the research community, and the general public. An aggressive outreach program has been undertaken to attract potential data providers and system users by: (1) providing tools and services to data providers, (2) building and maintaining an online data archive, and (3) demonstrating the utility of the system through new research. More specifically, a series of potential applications has been devised to help illustrate the utility of the OBIS-SEAMAP system for biogeographic research, resource management and marine conservation (Table 1). This thematic approach serves several purposes, including helping to define potential goals and products, prioritizing the raw data needs and tool development, providing a form of synthetic atlas to identify spatial data gaps for future research needs, and facilitating effective outreach and public engagement.

Biodiversity data network

The OBIS-SEAMAP program and the OBIS network, in general, are examples of a larger community of emerging ocean biodiversity informatics programs and activities (Costello & Vanden Berghe 2006, in this Theme Section). Ecoinformatics focuses on the development of technologies to enhance the discovery, exchange, and analysis of ecological data (see http://ecoinformatics.org). In order to facilitate the discovery and exchange of data between programs, the central OBIS facility has adopted the use of the Darwin Core protocol using XML (Extensible Markup Language) and DiGIR (Distributed Generic Information Retrieval: http://digir.sourceforge.net) as a standard Internet exchange language and database access package for search and retrieval of records between participating network data nodes. OBIS-SEAMAP is participating in the OBIS network through the use of DiGIR's client package, thus making its data available to the public through the central OBIS search interface. OBIS, in turn, is a data provider to the Global Biodiversity Information Facility (GBIF), which also uses DiGIR and XML.

Data Mines, Factories, and Pipelines

Development of the Internet-enabled OBIS-SEAMAP system requires flexible access tools for end users, rapid ingestion of data from providers, and on-the-fly server-to-server data conduits to and from institutional partners for overlay and analysis. These 3 components can be termed, respectively, *Data Mines*, *Factories*, and *Pipelines* (Fig. 1).

The *Data Mine* allows users to browse, search, map, and download archived biogeographic data. A custom search interface has been created, for taxonomic, attribute, and spatial searches. All datasets may be

Table 1. Potential applications illustrating the utility of the OBIS-SEAMAP (Ocean Biogeographic Information System–Spatial Ecological Analysis of Marine Megavertebrate Animal Populations) system for biogeographic research, resource management, and marine conservation

Theme	Rationale	Examples
Climatological setting	To illustrate some of the pervasive biogeographic patterns of species distribution in the world's ocean	Spatial gradients in community composition (e.g. onshore–offshore) Temporal changes in species distributions (e.g. seasonal migrations)
Anomalous conditions	To document temporal changes in communities	Interannual shifts in species distributions and community composition (e.g. El Niño–La Niña conditions)
Long-term change	To discriminate between anthropogenic impacts and natural variability in megavertebrate populations	Population trends in abundance due to anthropogenic impacts (e.g. bycatch and overexploitation) and shifts in population ranges (e.g. climate change)
Management of marine resources	To help delineate important habitats deserving protection and to determine national responsibilities for the management and conservation of protected species	Important habitats of protected species (e.g. migration corridors) Stock structure of megavertebrates (e.g. distributions, movements)
Conservation	To mitigate existing anthropogenic impacts and to identify additional threats to protected species	Overlap with potential impacts (e.g. oil and gas exploration, fisheries effort, and shipping lanes) Spatial and temporal areas where impacts take place (e.g. fisheries catch and bycatch)

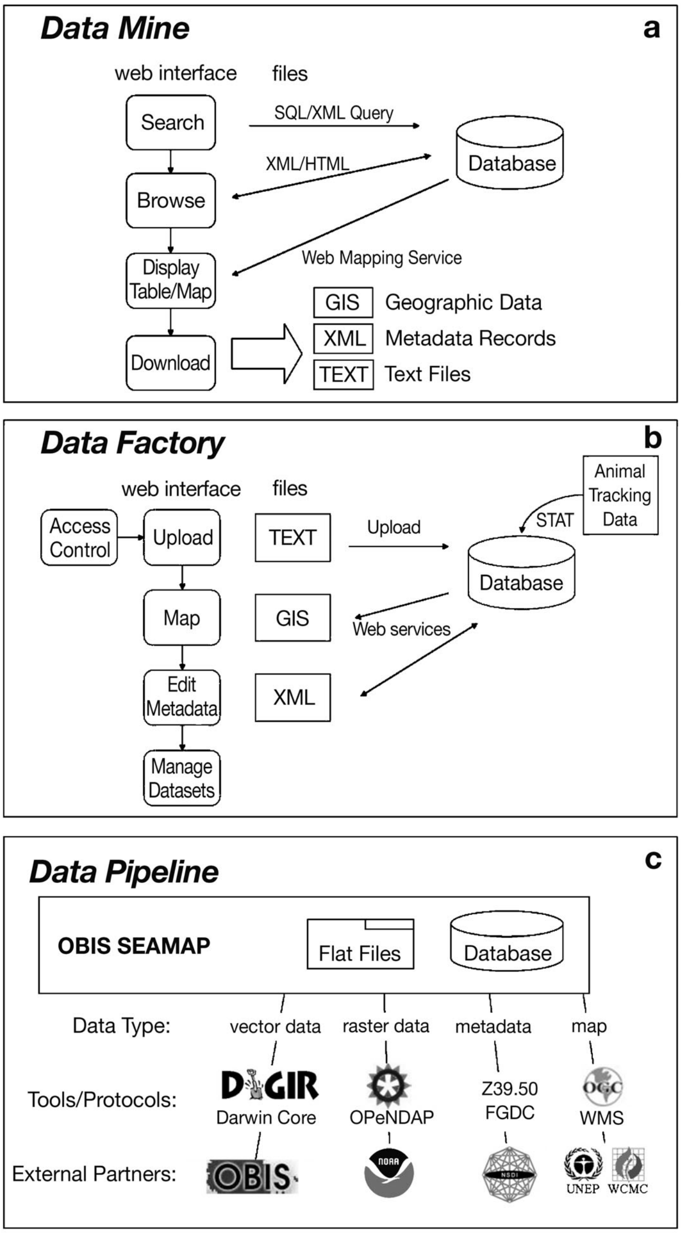


Fig. 1. Conceptualization of the OBIS-SEAMAP services: (a) *Data Mine* for end users, (b) *Data Factory* for data providers, and (c) *Data Pipeline* for server-to-server data exchange

browsed to a detail page, which provides summary information, a species list, FGDC-compliant (www.fgdc.gov) metadata, and original provider contact information. The species list for each dataset links to individual species profiles, which, in turn, contain links back to all relevant dataset detail pages. The online mapping component (described below) allows users to interactively select species, background environmental layers, and spatial extent.

The *Data Factory* enables data providers and managers to directly upload data into the OBIS-SEAMAP system, match taxonomic codes to species names, enter metadata, and 'publish' their data, making it available to the public. The most common format for data exchange are delimited text files, which can be output from most data-storing programs, and is the preferred format for uploading data. Spatial and temporal extents, along with full taxonomic hierarchies, are calculated for automated creation of FGDC-compliant metadata. A content management system (CMS), or more specifically Plone (http://plone.org), has proven extremely useful in supporting the transfer of all data-related files and content from data providers, as well as providing useful tools within our group for general project management and web content creation.

In addition to local data processing, the OBIS-SEAMAP system is capable of receiving and processing automated data uploads from the Argos satellite tracking system (Fig. 1b). Automated data upload and processing of satellite tracking provides sophisticated processing tools that are very attractive to potential data providers. The STAT program (Coyne & Godley 2005), developed by a member of the OBIS-SEAMAP team, provides a prototype of the full functionality that is currently being extended to the OBIS-SEAMAP system.

Finally, the *Data Pipelines* not only disseminate data to partners, such as to OBIS through DiGIR, but also consume data to create map overlays without storing oceanographic data locally. Using widely adopted protocols and tools, server-to-server Internet communication automates the distribution of information. For example, the OGC Web Mapping Service (WMS) protocol is now used by the OBIS-SEAMAP map server to retrieve date-specific oceanographic images of sea-surface temperature, sea heights, and winds on-the-fly from the NASA Jet Propulsion Laboratory's WMS server (http://seablade.jpl.nasa.gov/de.shtml). This allows end users to visualize biological data within an oceanographic context from the OBIS-SEAMAP mapping interface without the need to store a single byte (of the existing terabytes) of oceanographic information.

Flexible, taxonomic database

Long-term offsite storage and viewing of data specific to individual research projects provide limited benefits if users are forced to use an inflexible database design that does not appropriately reflect the original data. For this reason, OBIS-SEAMAP has adopted a flexible relational database design, whereby a single dataset summary table is linked to multiple individual dataset tables. Each dataset table has a subset of common fields with any number of additional fields allowed. Data for common fields across all datasets are easily viewed using a database query, while data specific to an individual dataset are maintained within their own table. The flexibility inherent in this database design provides a complete picture of the entire original dataset, which is further appealing for data providers as an offsite backup option.

Taxonomic data storage presents another interesting challenge, as taxonomic names and hierarchies can change over time. The Integrated Taxonomic Information Service (ITIS) (www.itis.usda.gov) and the partnered Species 2000 Catalogue of Life Programme (www.sp2000.org) represent efforts to provide standard taxonomic nomenclature through web services. OBIS-SEAMAP uses the ITIS program's XML service to match common and scientific names with a taxonomic serial number (TSN). All information, including vernacular names and parent and child taxa can be accessed using this XML service. In the OBIS-SEAMAP, 'Data Factory' scripts automatically match recorded species names to TSNs, while providing an easy-to-use interface for the provider to search and match any remaining TSNs after a dataset is uploaded. Once the TSN has been obtained, all related information is readily accessible for populating the database and for linking to taxonomic information. For example, the metadata record for each dataset lists the full taxonomic hierarchy, including rank, scientific name, and common name, of all species found in the dataset, in conformance with the FGDC Biological Profile. This task would be extremely time consuming if not for such centralized taxonomic services. Having information taxonomically indexed is especially important for retrieving all query results at higher taxonomic rankings (e.g. phylum, class, order, genus) beyond just the species level.

Internet mapping

Interactive visualization of geospatial data is now feasible with existing Internet Mapping Services (IMS). The OBIS-SEAMAP team is constantly evaluating emerging open-source and commercial Internet mapping software to select the most robust open-standards applications to meet the needs of the marine user community. The OBIS-SEAMAP mapping interface currently uses a PostgreSQL database, a PostGIS geo-database connector, and the University of Minnesota's MapServer on a Linux operating system platform. This current configuration provides a robust, open-source, and open-standards tool kit that can be emulated by other global or regional information system projects. The open-standards applications allow for the sharing of data and imagery through web services conforming to OGC and OBIS standards.

While the central OBIS portal facility maintains a mapping interface, OBIS-SEAMAP data providers specifically requested the ability to plot sampling effort and vessel trackline information along with their animal observation data. It is often equally useful to know where animals were not observed as it is to know where they were observed. Many important statistical analyses and ecological models require the explicit sampling effort data (e.g. ship and aircraft track line data) and condition (e.g. Beaufort sea state) for proper calculation. For aerial and boat surveys, data providers requested that a line of cruise effort be shown, and for satellite-tracked animals, an inferred track between known locations.

The OBIS-SEAMAP system currently accommodates sampling effort and telemetry tracking data types. In order to support this type of data across the entire OBIS network, the OBIS schema will need to be expanded. The inclusion of survey effort data will require the identification of beginning and ending coordinates for ship or aircraft search effort tracklines and survey conditions. In addition, the inclusion of telemetry tracking data will require (animal) series identifiers, as well as position quality codes for the interpretation of satellite position data. Representation of tracking data from archival data collectors will also require schema modifications in order to accommodate subsurface, 3-dimensional dive data collected for the animals. All of these additions to the general schema are currently under consideration to allow more general use and exchange of these classes of data. By moving beyond the display of simple point data, OBIS-SEAMAP encourages the CoML community to provide the necessary standardized effort data required to estimate the distribution and abundance of marine animals.

Future technology trends

The further development of web service architectures based on common standards (e.g. OPenDAP, OGC, SOAP, Marine XML, Digital Object Identifiers etc.) will allow for increased interoperability between the marine biogeographic observation community and the marine oceanographic observation community. The development of modular scripts for scientific workflow modeling programs (e.g. Kepler, JPL SciFlo, ESRI ModelBuilder etc.) and semantics to intelligently exchange across data schemas will facilitate easier, more comprehensive analysis of marine animals and their dynamic environments.

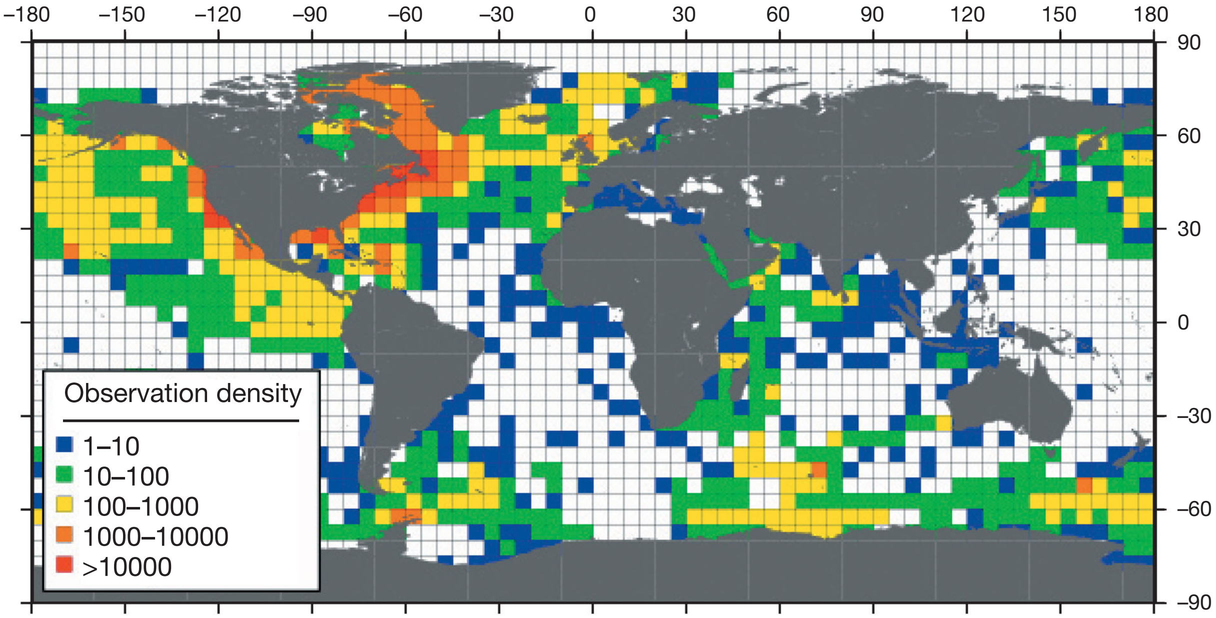


Fig. 2. Spatial density of available marine mammal, sea turtle, and sea bird observations contained in the OBIS-SEAMAP system as of September 20, 2005 in 5 × 5 degree cells (n = 1 144 248 observations)

CONCLUSIONS

The creation of a seamless biogeographic information system, integrated with physical (e.g. ocean physics) and biological (e.g. ocean productivity) datasets, will be the main legacy of the OBIS program (Grassle & Stocks 1999, Zhang & Grassle 2002). This interoperable system will place a wealth of physiographic (e.g. habitats) and biogeographic (e.g. species distribution and abundance) data for a broad array of marine organisms and areas of the globe at the fingertips of researchers, students, managers, and policy decision-makers worldwide. This global, interdisciplinary, multi-taxa perspective will provide insights into spatial and temporal changes in ocean productivity, biogeographic patterns, and ecosystem structure around the globe.

The OBIS-SEAMAP digital database of marine mammal, seabird, and sea turtle observations is available online and is fully accessible at: http://seamap.env.duke.edu. At the time of writing (September 2005), the database includes 163 datasets, comprising >1 million records, spanning from 1935 to 2005 (Table 2). The website also includes >100 species profiles and other background and outreach materials. The global distribution of observation data is depicted in Fig. 2. A web mapping interface allows the interactive display, query, and analysis of this database through the OBIS-SEAMAP homepage.

The diverse array of OBIS-SEAMAP products are already providing marine biogeographers and resource managers with essential information for the study and conservation of marine mammal, bird, and turtle species (e.g. MPA News, March 2004, p. 6). In particular, the web-based query, subset, and data-export tools, and the extensive supporting documentation (e.g. survey methods, ancillary environmental information) have clearly enhanced the utility of the database for scientific research.

OBIS-SEAMAP is also engaging the general public (e.g. educators, students, national governmental organizations) by providing web-based mapping tools to display survey data in conjunction with environmental information and to summarize temporal and spatial patterns of species occurrence in an easily digestible format (e.g. The Society for Conservation Geographic Information Systems e-Newsletter, January 2003, p 5). This outreach is enhancing the public's appreciation for marine ecology and biogeographic patterns by pro-

Table 2. Observation and sampling effort data contained in the OBIS-SEAMAP data archive listed by survey type (rows) and taxa (columns)

Survey platform	Number of observations Birds	 Mammals	 Turtles	Total observations	Number of datasets
Boat	271 909	80 374	1 953	354 236	93
Plane	503 605	110 649	4 866	619 120	57
Shore	127 759	2 112	–	129 871	4
Tag	8 797	22 637	9 587	41 021	9
Total	912 070	215 772	16 406	1 144 248	163

viding a database of supporting ecology and natural history information (e.g. species profiles, range maps, and dietary information) necessary to interpret the web-based maps and database summaries.

A clear sign of the continued success of the OBIS-SEAMAP program will be the increasing participation, collaboration, and data sharing between researchers, managers, and educators around the globe. Progress has already been made with the launch of several meta-analysis projects using or building upon the OBIS-SEAMAP framework. A prime example of the direct application of the OBIS-SEAMAP datasets involves new analyses incorporating data and expertise from a number of OBIS-SEAMAP data providers to predict marine mammal habitats, a project that is supported by the Strategic Environmental Research and Development Program (SERDP).

This research will use spatial analysis techniques to assist the U.S. Navy in planning military readiness exercises in U.S. coastal waters. A similar project, built upon the OBIS-SEAMAP framework, is assessing the global status of sea turtles by looking at spatial and temporal trends in nesting data (Hutchinson et al. 2005). This project, which not only uses OBIS-SEAMAP data, but also feeds them back into the system, was a major impetus for the addition of new time-series visualization and analysis tools within OBIS-SEAMAP. Both of the marine mammal and sea turtle application projects described above have, in large part, been made possible through the development of the OBIS-SEAMAP data commons.

On the basis of our experience, we contend that the development of large biogeographic data commons will benefit ecological research, not only by compiling the vast datasets required to ask large-scale climatic and conservation questions, but also by acting as a catalyst for the development of the collegiate and collaborative community atmosphere necessary to undertake the large multi-investigator studies required to answer these pressing ecological questions.

Acknowledgements. We thank M. Costello and E. Vanden Berghe for assembling this 'Ocean Biodiversity Informatics' theme section. We are also indebted to the OBIS-SEAMAP steering committee (S. Batten, D. Boersma, J. Croxall, P. Dayton) and the members of the OBIS portal and OBIS International Committee for advice and support. We particularly pay tribute to all of the data providers, especially from the NOAA National Marine Fisheries Services, who have attended the annual OBIS-SEAMAP data provider meetings and contributed large datasets, as well as significant guidance. Funding for this work has been provided by the Alfred P. Sloan Foundation, the National Oceanographic Partnership Program, the U.S. Office of Naval Research, the Census of Marine Life, and the Ocean Biogeographic Information System. We also recognize the oceanographic community for its open-access ethos that has made this work possible.

LITERATURE CITED

Alldredge AL, Bradley DL, Butterworth D, Steele JH (1999) Assessing the global distribution and abundance of marine life: summary of a workshop sponsored by the Sloan Foundation and the Office of Naval Research, January 13–15, Monterey, California. Oceanography 12(3): 41–46

Arzberger P, Schroeder P, Beaulieu A, Bowker G and 5 others (2004) An international framework to promote access to data. Science 303:1777–1778

Ausubel JH (1999) Toward a census of marine life. Oceanography 12(3):4–5

Block BA, Costa DP, Boehlert GW, Kochevar RE (2003) Revealing pelagic habitat use: the tagging of Pacific pelagics program. Oceanol Acta 25:255–266

Bradley DL (1999) Assessing the global distribution and abundance of marine organisms. Oceanography 12:19–20

Costello MJ, Vanden Berghe E (2006) 'Ocean biodiversity informatics: a new era in marine biology research and management. Mar Ecol Prog Ser 316:203–214

Coyne MS, Godley BJ (2005) Satellite Tracking and Analysis Tool (STAT); an integrated system for archiving, analyzing and mapping animal tracking data. Mar Ecol Prog Ser 301:1–7

Decker CJ, O'Dor R (2002) A census of marine life: Unknowable or just unknown? Oceanol Acta 25:179–186

Grassle JF, Stocks KI (1999) A global ocean biogeographic information system (OBIS) for the census of marine life. Oceanography 12(3):12–14

Hutchinson BJ, Mast RB, Pilcher NJ, Seminoff JA (2005) Marine turtle specialist group news: overview of activities for a new year. Mar Turtle Newsl 108:15–17

Levi C, Stone G, Schubel JR (1999) Censusing non-fish nekton. Oceanography 12(3):15–18

McGowan JA (1990) Climate and change in oceanic ecosystems: the value of time series data. Trends Ecol Evol 5(9):293–300

NRC (National Research Council) (1996) Understanding marine biodiversity. National Academy Press, Washington, DC

O'Dor RK (2003) The unknown ocean: the baseline report of the census of marine life research program. Consortium for Oceanographic Research and Education, Washington, DC

Pierott-Bults AC (1997) Biological diversity in oceanic macrozooplankton: more than just counting species. In: Ormond RFG, Gage JD, Angel MV (eds) Marine biodiversity: patterns and process. Cambridge University Press, Cambridge, p 69–93

Stone G, Schubel J, Tausig H (1999) Electronic marine animal tagging: new frontier in ocean science. Oceanography 12(3):24–27

Welch DA, Boehlert GW, Ward BR (2003) POST—the Pacific Ocean salmon tracking project. Oceanol Acta 25:243–253

Zhang YQ, Grassle JF (2002) A portal for the ocean biogeographic information system. Oceanol Acta 25:193–197

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: September 27, 2005; Accepted: December 2, 2005
Proofs received from author(s): May 26, 2006

Vol. 316: 247–255, 2006

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

Continuous Plankton Recorder database: evolution, current uses and future directions

Darren Stevens[1,*], Anthony J. Richardson[1,2,3], Philip C. Reid[1]

[1]Sir Alister Hardy Foundation for Ocean Science (SAHFOS), The Laboratory, Citadel Hill, The Hoe, Plymouth PL1 2PB, UK
[2]CSIRO Marine Research, Cleveland, Queensland 4163, Australia
[3]Department of Applied Mathematics, University of Queensland, St Lucia, Queensland 4072, Australia

ABSTRACT: The Continuous Plankton Recorder (CPR) survey, operated by the Sir Alister Hardy Foundation for Ocean Science (SAHFOS), is the largest plankton monitoring programme in the world and has spanned >70 yr. The dataset contains information from ~200 000 samples, with over 2.3 million records of individual taxa. Here we outline the evolution of the CPR database through changes in technology, and how this has increased data access. Recent high-impact publications and the expanded role of CPR data in marine management demonstrate the usefulness of the dataset. We argue that solely supplying data to the research community is not sufficient in the current research climate; to promote wider use, additional tools need to be developed to provide visual representation and summary statistics. We outline 2 software visualisation tools, SAHFOS WinCPR and the digital CPR Atlas, which provide access to CPR data for both researchers and non-plankton specialists. We also describe future directions of the database, data policy and the development of visualisation tools. We believe that the approach at SAHFOS to increase data accessibility and provide new visualisation tools has enhanced awareness of the data and led to the financial security of the organisation; it also provides a good model of how long-term monitoring programmes can evolve to help secure their future.

KEY WORDS: Data accessibility · Visualisation tools · CPR data

INTRODUCTION

Large data repositories are common within the physical, chemical and biological oceanographic community. It is generally accepted that such data should be shared openly and freely, providing a wealth of valuable information for researchers, ecosystem modellers, policy makers and the general public. However, data are not always easily accessible, and the volume of diverse information can be difficult to integrate and synthesise (Vanden Berghe et al. 2004). This requires organisations to continually make data more accessible and provide simple yet evocative visual representations of data in easily digestible forms. Here we describe these challenges as they pertain to the largest plankton monitoring programme in the world, the Continuous Plankton Recorder (CPR) survey.

The CPR survey provides a long-term baseline of the near-surface distribution, abundance and diversity of phyto- and zooplankton. This has been used to assess biodiversity (Beaugrand et al. 2000, Beaugrand & Ibañez 2002), especially in terms of impacts of climate change (Beaugrand et al. 2002, Richardson & Schoeman 2004), over-fishing (Reid et al. 2000), pollution (Batten et al. 1998), eutrophication (Edwards et al. 2001a) and the spread of invasive species (Edwards et al. 2001b). The survey was initiated in 1931 and has been operated since 1991 by an international charity, the Sir Alister Hardy Foundation for Ocean Science (SAHFOS).

The CPR itself is a high-speed plankton sampler towed monthly behind commercially operated ships of opportunity. Since its inauguration the survey has operated in the North Sea and, since 1939, in the North

*Email: dpst@sahfos.ac.uk

Atlantic, with a break during World War II. Operations started in the North Pacific in 1997. Water enters the CPR and flows down a tunnel through a silk filtering mesh. Upon return to the laboratory the silk is cut into 10 nautical mile sections that equate to samples (for more information see Warner & Hays 1994, Richardson et al. 2006).

In this contribution we describe the CPR database and the approach taken at SAHFOS over the years to increase data accessibility and provide new visualisation tools. We first describe the evolution of the CPR database and access issues. We argue that as the data have become more accessible, both to researchers within and outside SAHFOS, not only is more high-quality research being conducted, but the CPR data are being used more directly in marine ecosystem management. We then describe 2 recent visualisation tools, the WinCPR North Sea plankton browser and the digital CPR Atlas, which make CPR data available to both researchers and non-plankton specialists. Enhanced awareness and use of CPR data have contributed to the financial security of the organisation, and provide a model of how other long-term monitoring programmes can evolve to help secure their future.

CPR DATA

Since September 1931, the CPR has been towed >5 million nautical miles, with 196 120 samples counted (see Fig. 3). This equates to >2.3 million records of individual taxa being present or ~90 million data points, including zero records. Samples are analysed for plankton abundance under a microscope in 3 stages: (1) phytoplankton, (2) small zooplankton <2 mm counted in a traverse across the silk and (3) zooplankton >2 mm that are removed from the silk and counted under low magnification. More than 450 different taxa have been identified, over half to species level (Reid et al. 2003). A description of the methods of the CPR survey, along with a list of all taxa and relevant information for each taxon, is given in Richardson et al. (2006).

In addition to plankton data, concurrent environmental data, such as temperature, chlorophyll and salinity, are measured. Environmental data are collected on approximately half of the CPR routes; this will hopefully be expanded to all CPR deployments in the future, as funding allows. There is also extensive auxiliary information on the attributes of each tow and sample. In terms of the tow, the name and average speed of the ship, the latitude and longitude of deployment, retrieval and course changes during the passage of the ship, the identification numbers of both the CPR and the interchangeable internal mechanism and the propeller angle of the CPR are recorded. For each sample, the location and local time at the midpoint of the course is calculated, and the name of the person who counted the sample is documented.

STORAGE AND ACCESS

Over the last 70 yr, the storage methods for CPR data have changed as technology has advanced (Stevens & Reid 2004), allowing improved data access (Fig. 1). Prior to 1969, CPR data were stored on cards and large maps, with data in >1 format (e.g. by tow and by species). As all calculations were carried out by hand, analysis of the data was time consuming and limited to basic operations. In 1969, the first computerised database containing CPR data was developed on a KDF9 computer. Initially this database only stored processed data (monthly means for standard areas), but no raw data, and access by researchers was limited because the database was housed offsite. By the early 1970s, a database to store raw data from newly analysed samples was developed, and the CPR team took steps to ensure that historical data were entered retrospectively. Data were now accessible faster than before, but the employment of a programming specialist was necessary to extract data from the file-based database. The database has evolved considerably over the years; initially it was accessed using ALGOL (ALGOrithmic Language), then by IMP (Implementation Language), after that concurrently by PASCAL and FORTRAN (IV) G, and then by FORTRAN 77 running on an IBM OS/2 platform. In 1993 the first relational database for CPR data was developed in ORACLE, but this was never fully adopted because of financial constraints.

In 1995, the CPR database was transposed into the ACCESS relational database still in use today. This gave researchers within SAHFOS easier access to raw data. However, calculations for sample position and time were still processed by the old FORTRAN system. A further difficulty in gaining access to the data was the limited number of staff members with the skills required. Almost all research at this time was carried out by those directly involved with the survey, as data were not easily available to external researchers.

By the end of the 20th century a significant change in the philosophy at SAHFOS concerning data accessibility had evolved. In May 1999, SAHFOS amended its data policy to comply with the emerging Global Ocean Observing System (GOOS) programme, making the data freely available for non-profit research (Reid et al. 2003). Since then, the CPR survey has formed part of the Initial Observing Programme of GOOS. As part of our commitment to GOOS, data on important indicators of primary (phytoplankton colour) and secondary

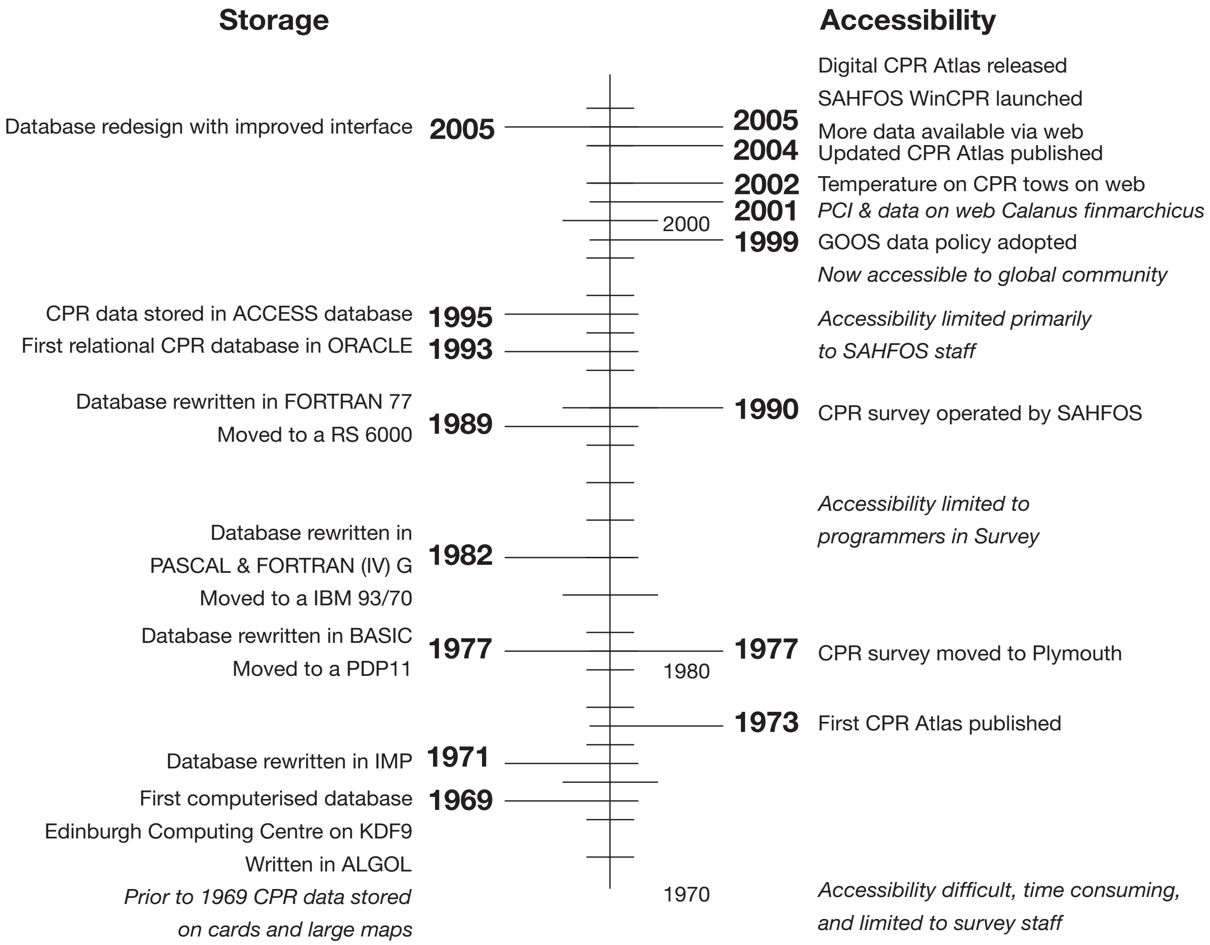

Storage
Accessibility
Database redesign with improved interface 2005
CPR data stored in ACCESS database 1995
First relational CPR database in ORACLE 1993
Database rewritten in FORTRAN 77
Moved to a RS 6000 1989
Database rewritten in
PASCAL & FORTRAN (IV) G 1982
Moved to a IBM 93/70
Database rewritten in BASIC 1977
Moved to a PDP11
Database rewritten in IMP 1971
First computerised database 1969
Edinburgh Computing Centre on KDF9
Written in ALGOL
Prior to 1969 CPR data stored
on cards and large maps
2000
1980
1970
Digital CPR Atlas released
SAHFOS WinCPR launched
2005 More data available via web
2004 Updated CPR Atlas published
2002 Temperature on CPR tows on web
2001 PCI & data on web Calanus finmarchicus
1999 GOOS data policy adopted
Now accessible to global community
Accessibility limited primarily
to SAHFOS staff
1990 CPR survey operated by SAHFOS
Accessibility limited to
programmers in Survey
1977 CPR survey moved to Plymouth
1973 First CPR Atlas published
Accessibility difficult, time consuming,
and limited to survey staff

Fig. 1. Changes in storage and accessibility of Continuous Plankton Recorder (CPR) data through time. GOOS: Global Ocean Observing System; PCI: phytoplankton colour index; SAHFOS: Sir Alister Hardy Foundation for Ocean Science

(*Calanus finmarchicus*) productivity are freely available directly from the SAHFOS website (www.sahfos.org). Data are available as pre-processed monthly means for CPR standard areas (pre-defined areas used historically within the survey). At present, no information is collected on who is accessing these data or the number of times they are being downloaded; this should be redressed in the future.

Currently, researchers requiring any data other than phytoplankton colour and *Calanus finmarchicus* abundance in CPR standard areas need to complete, sign and return a data licensing agreement available from the SAHFOS website. A dedicated full-time database manager has processed these data requests since August 1999, enabling us to monitor the increase in data requests (Fig. 2). We believe this is linked to the increased scientific profile of SAHFOS, enhancing awareness of CPR data, and the more open data policy, as well as the investment in computer hardware and software improving accessibility. Since this time, there have been >100 requests for CPR data from 14 different countries across 3 continents (Canada, France, Germany, Iceland, Italy, Netherlands, Norway, Portugal, Republic of Ireland, Spain, South Africa, UK and USA), emphasising the research and management value of the archive.

In 2001 the FORTRAN programs were rewritten in Visual BASIC for ACCESS, because of concerns about future compatibility. In 2002, temperature data along selected CPR routes since 1996 were made available via the SAHFOS website.

During 2004, in collaboration with the Ocean Biogeographic Information System (OBIS) project of the Census of Marine Life (CoML), CPR plankton presence data have been made available via the World Wide Web (www.iobis.org). Consequently, via OBIS, CPR data are available through the Global Biodiversity Information Facility (GBIF). The CPR dataset supplies more records to OBIS than any other provider and is thus also a significant provider to GBIF. OBIS allows users to draw simple distribution maps for nearly 40 000 marine species, from sponges to whales. Making CPR

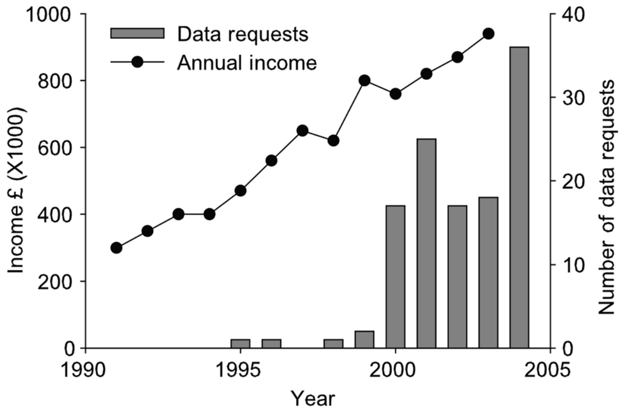


Fig. 2. Bar chart showing the number of external data requests, with indication of SAHFOS income through time

plankton data accessible via such large international data portals will further enhance the awareness, accessibility and use of the dataset.

The policy currently adopted by SAHFOS provides data as monthly and annual means. Researchers requiring access to raw sample data need to visit SAHFOS in Plymouth (UK) to obtain the data. This allows the researcher to witness the process of counting samples and become familiar with the idiosyncrasies of the CPR methodology and data.

PRODUCTS

Recent efforts have focused on making CPR data more available by developing software to allow users easier access and increased flexibility of data interrogation. Two such products have been developed, i.e. SAHFOS WinCPR and the digital CPR Atlas. These products provide oceanic researchers with graphical outputs that aid data interpretation, increasing the accessibility of CPR data to a wider audience. Both WinCPR and the CPR Atlas were originally programmed in MATLAB, to overcome the problem of spatial and temporal biases in the sampling. The aim was to provide simple access to CPR data in a summary form, but this was hindered by the sophistication and cost of the software. Therefore, user-friendly, WINDOWS-compatible front ends to the browsers were built.

WinCPR is a gridded database browser of North Sea plankton, containing data from a 50 yr period (1948 to 1997). It targets not only the marine science community, but a wider audience, including the general public and students from schools and universities (Vezzulli & Reid 2003, Vezzulli et al. 2004). The user-friendly information and buttons on the opening page are visually appealing, clear and allow the user to perform sophisticated analysis quickly. The software allows users to produce output summarising changes in monthly and interannual abundance of plankton taxa. The grid consists of 172 pixels centred on 1° longitude × 0.5° latitude. A total of 110 plankton taxa, as well as phytoplankton colour (an estimate of phytoplankton biomass), have been averaged for each month within a 40 or 50 yr time span (50 yr for zooplankton and phytoplankton colour, 1948 to 1997; 40 yr for phytoplankton, 1958 to 1997). Products available include distribution maps averaged annually or over the entire period, graphs of abundance through time and seasonal cycles, and month-by-year contour plots for individual and groups of pixels. Underlying gridded data can also be exported; this will be particularly useful for ecosystem modellers in validation and initialisation. The software is available for download via the SAHFOS website (www.sahfos.org/winCPR.htm).

The other major product recently developed to increase data accessibility is the digital CPR Atlas. The first hardcopy CPR Atlas on spatial distribution of plankton in the North Atlantic was published in 1973, and was based on only ~40 000 CPR samples from 1958 to 1968 (see Fig. 3 for number of samples per grid square). This atlas contributed to our knowledge of the biogeography of ~260 taxa.

Work was started on updating the atlas in 2001, for the celebration of the 70th anniversary of the CPR survey, and was published as a supplement of the Marine Ecology Progress Series in 2004 (CPR Survey Team 2004). This new hardcopy atlas is based on >150 000 CPR samples, collected from 1958 to 1999. Fig. 3 shows the distribution and number of samples in the North Atlantic, with highest numbers in the NE Atlantic and North Sea. A statistically robust procedure was developed to reduce bias associated with the irregular sampling through time of the CPR survey (Beaugrand 2004). Maps were produced using Lambert conical projection rather than the Mercator projection used in the 1973 edition, removing the problem of distortion for large areas away from the Equator. This atlas will be a powerful tool for researchers, an invaluable tool for para-taxonomists needing to know whether a species is found in a certain area, but will also help to define important biological regions for international biodiversity initiatives such as the CoML.

The new digital CPR Atlas is based on the hardcopy atlas released in 2004, but includes several powerful added features. It not only allows users to view distribution maps of the North Atlantic for the entire period (1958 to 1999), but also enables maps to be drawn for each decade, and for day and night periods. Decadal maps will help document the spread of non-indigenous species and provide a critical baseline to assess climate impacts on plankton. For example, it is clear from the

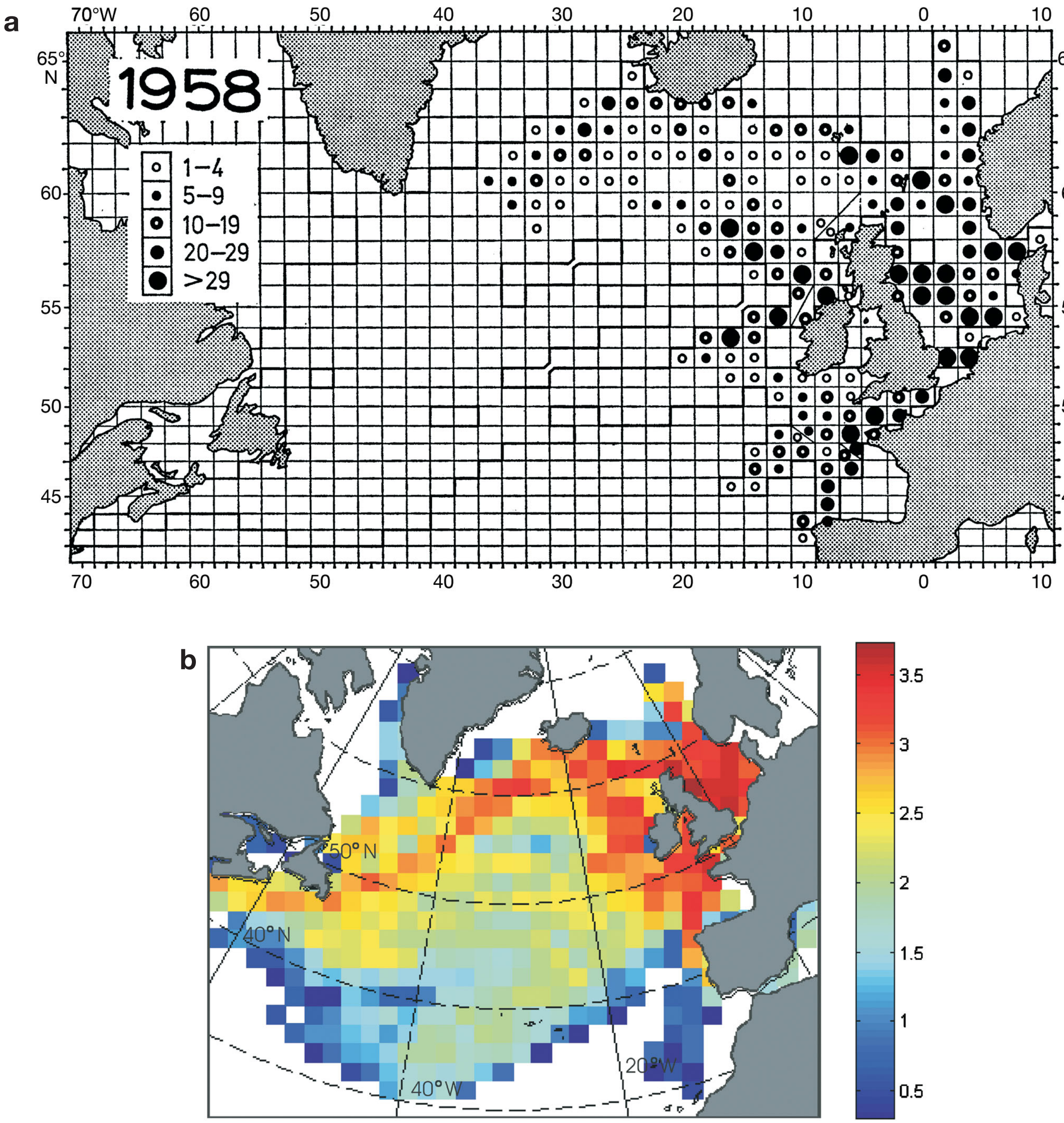


Fig. 3. Number and distribution of samples throughout the North Atlantic and North Sea for the CPR Atlas in: (a) 1958 (1973 version) and (b) 1958 to 1999 (2004 version; note: scale here is logarithmic base 10)

decadal maps produced that a number of subtropical species have moved further north over the last 4 decades as sea temperatures have warmed (also see Beaugrand et al. 2002). Maps showing day–night differences highlight species that undergo extensive diel vertical migration (also see Beaugrand et al. 2001). The digital CPR Atlas will become available in 2006.

USAGE

Research

CPR data have underpinned high-quality research for >70 yr, and the growth in resulting publications has continued to increase (Fig. 4). A comprehensive bibli-

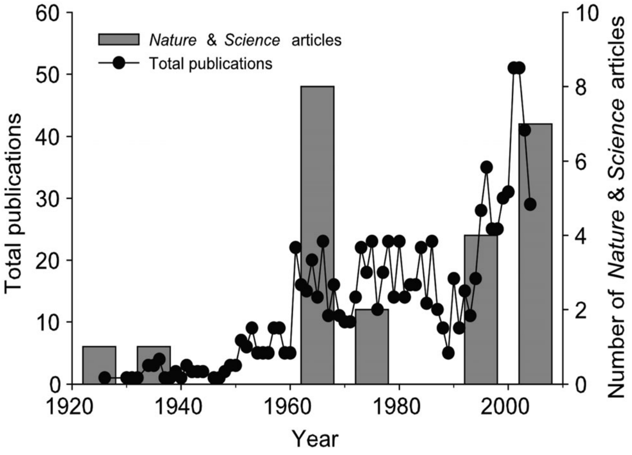


Fig. 4. Total number of publications annually using CPR data (refereed and grey literature) and the number of articles in *Nature and Science* by decade

ography of publications relating to the CPR survey is available on the SAHFOS website, managed by the National Marine Biological Library, Plymouth, UK.

In the early years, CPR data were used to describe and document general plankton ecology in terms of the timing of phyto- and zooplankton blooms (Colebrook & Robinson 1961) and regional differences in productivity (Colebrook 1964). At this time data were used almost exclusively by researchers within the CPR survey. More recently, the long time series has been used as a baseline against which global change can be assessed. Much of the research has focused on impacts of our changing climate on biodiversity (Beaugrand et al. 2000, 2002), plankton productivity (Richardson & Schoeman 2004), timing of seasonal cycles and synchrony between successive trophic levels (Edwards & Richardson 2004) and fisheries productivity (Beaugrand et al. 2003, Beaugrand & Reid 2003). Other impacts of global change assessed include over-fishing (Reid et al. 2000), pollution (Batten et al. 1998), eutrophication (Edwards et al. 2001a) and the spread of invasive species (Edwards et al. 2001b). With greater data accessibility over recent years, researchers not affiliated with SAHFOS have been increasingly using CPR data. This is highlighted by the growth in external data requests since records began in 1995 (Fig. 2) and a number of recent high-profile publications driven externally to SAHFOS (e.g. DeYoung et al. 2004, Thompson et al. 2004).

Marine management

Information from the CPR survey has been used extensively to support marine management in the areas of fisheries and, environmental protection, and in the study of ecosystem response to environmental change (Brander et al. 2003). In fact, the survey was initiated to better manage the herring fishery in the North Sea by reducing its variability and improving its efficiency (Reid et al. 2003). In contemporary times, the survey has focused on documenting changes in ecological indicators that are sensitive to alterations in ecosystem health. CPR data were used in the Quality Status Report of the North Sea (e.g. North Sea Task Force, NSTF 1993) and in an annual Ecological Status Report of the North Atlantic (Edwards et al. 2004; see www.sahfos.org). Specific procedures have also been designed and implemented for monitoring climate-driven changes in copepod biodiversity (Beaugrand 2004, Beaugrand & Ibañez 2004). Other indicators sensitive to environmental change that have been developed include the presence and location of harmful algal blooms, the relative dominance of mero- and holozooplankton, unusual range extensions and the dominance of warm- or cold-water taxa. Such indicators contribute to assessments of the health of European waters by the UK Department of Environment, Food and Rural Affairs (Reid et al. 2004), the UK Joint Nature Conservation Committee (e.g. Edwards & John 1998), the European Environment Agency, the International Council for the Exploration of the Sea and OSPAR (Reid 1999; Quality Status Report 2000).

THE FUTURE

SAHFOS is actively enhancing methods of data storage and developing new ways of making CPR data more accessible.

WinCPR

We have acquired funding to further develop the WinCPR software through 2 future expansions. The first of these is to extend the dataset for the North Sea up to the present, so the database will include a further 6 yr of CPR data. There will also be an improved method of gridding, similar to the one used to create the CPR Atlas. The size of each pixel will be changed in an effort to reduce artefacts of the gridding process. This development will also include other environmental variables, such as sea-surface temperature (SST), cloud cover and wind speed, so that relationships between the environment and plankton can be assessed. The second phase of development is to extend the geographical coverage of the browser to the NE Atlantic.

Upgrade of the CPR database

The size of the CPR dataset makes data extraction sluggish under the current Microsoft ACCESS system. Investigations are underway to upgrade the system, with advice from the British Oceanographic Data Centre and the University of Plymouth. The new computerised database will include further variables such as the microscope number (important to assess whether there are inter-microscope differences), the height of the tow point, the length of the tow wire, location of stored sample in the sample archive, whether a sample was collected during the day or night, a link to a taxonomic manual that is in preparation (containing information on each species) and CPR data, currently in paper form, from 1931 to 1938. The new database will also allow improved quality control by providing information on the distribution of each taxon, to highlight potential irregularities, and by comparing samples with others taken along the same route. Easy access to the CPR samples within the archive via the database will allow taxonomic (Lindley 1982), genetic (Kirby & Reid 2001) and pollution (e.g. plastics, Thompson et al. 2004) research to be conducted retrospectively.

Database developers are presently taking a fresh look at the way CPR data are managed, from the moment the samples are returned to the laboratory to the point when standard analysis is complete and samples are archived. The aim of this project is not only to make the data more secure and easier to manage, but also to decrease the time between the arrival of samples at the laboratory and the release of data for use by researchers. CPR data will be entered directly into the computerised database, rather than entering plankton abundance counts on paper first, as is the current procedure. Another development will be the release of the phytoplankton colour index data earlier than the plankton abundance data. The new system will allow for phytoplankton colour data to be released within 6 to 8 wk of collection, in contrast to the current situation in which all plankton data are released in September of the following year.

Web access

Eventually we hope to make all CPR data available via the web. Web pages that allow data extraction need to include a level of security to ensure the integrity of the database and provide a record of data requests. The need for this streamlined method of data access is clear from the recent increase in data requests. Web access of CPR data will automate the process of data requests, providing monthly and annual means, along with raw data. To assist data downloads, summary statistics, such as the number of samples by month in an area of interest defined by the user, should be available interactively, prior to extracting plankton counts.

The plan is to add data on ecological indicators that have been identified in the Ecological Status Report available on the web. Initially this will be data on total copepods and SST from the Hadley Centre, The Met Office, UK, for CPR standard areas, to complement the phytoplankton colour data already available. Other data that should be included are the ratio of *Calanus finmarchicus* and *C. helgolandicus* abundance in the North Sea, the plankton phenology index for the central North Sea, the relative dominance of mero- and holozooplankton and the indicators of calanoid biodiversity.

Metadata are needed to ensure that researchers have a clear understanding of the data provided via the web. For example, from the 1960s to the 1980s Euphausiacea were divided into juveniles and adults, in addition to the total Euphausiacea category that has been counted consistently since 1948 (Southward et al. 2004). Recently, information describing each taxonomic entity in the CPR database has been published (Richardson et al. 2006) and should be incorporated into the SAHFOS website and CPR database.

Environmental data

Environmental information should be derived for each CPR sample and stored in the database. Data on SST, cloudiness and winds are available from the International Comprehensive Ocean–Atmosphere Data Set (www.cdc.noaa.gov/coads/) on a 1° monthly scale and dating back to 1860. Temperature, salinity and fluorescence data collected via CPRs will also be incorporated.

CONCLUSIONS

The CPR dataset is one of the most valuable marine biological surveys in existence. The CPR database has evolved with changes in technology, increasing data access to researchers within and outside SAHFOS. The organisation has found that it is imperative to plan for software development and budget for further design to ensure that the product will reach the widest audience. We have seen, over the last 10 yr, that efficient access to data has resulted in its increased use, in many high-impact publications within and outside the organisation, and in expanding the role it plays in marine management, all of which have raised awareness of the dataset. Monitoring the number of people accessing data and their affiliations provides a measure of the value of the dataset and helps justify continued financial support of the organisation into the future. SAHFOS has found that quick and easy access to CPR data will ensure the continued survival of the survey.

With the wealth of data now available on the web, it is easy for researchers to become overwhelmed. We have shown that to promote wider use, additional tools need to be developed to provide representations that are visually appealing. Two such software tools developed at SAHFOS, WinCPR and the digital CPR Atlas, allow easy access to CPR data for both researchers and non-plankton specialists. A critical feature of these tools is that they provide researchers with the ability to generate summary statistics instantly about the dataset, in order to refine their data queries. We believe that the approach at SAHFOS to increase data accessibility and provide new visualisation tools has enhanced awareness of the data and led to the financial security of the organisation; it also provides a good model of how long-term monitoring programmes can evolve to help secure their future.

The next challenge is linking the CPR database with other larger environmental and biological datasets. The role of distributed database systems, such as OBIS, will allow the integration of data from different sources providing greater spatial and temporal resolution. Data providers will be able to utilise the centralised tools provided by such systems, reducing duplication of development. Such initiatives will undoubtedly enable rapid data access, allowing more effective planning and targeted research. They will also help to ensure the continued survival of long-term monitoring programmes by developing new ways of making data accessible, informative and useful by a broad cross-section of the research community, policy makers and the general public.

Acknowledgements. The authors are grateful to all past and present members and supporters of SAHFOS and its predecessors; their sustained help has allowed the establishment and maintenance of the long-term CPR dataset. Currently, the main support for SAHFOS is from the Department of Environment, Food and Rural Affairs (UK) and the Natural Environmental Research Council (UK). Consortium support for the CPR survey is also provided by agencies from the following countries: UK, USA, Canada, Faroe Islands, France, Iceland, Ireland, Netherlands, Portugal, Spain, and by the International Oceanographic Commission and the European Union. We thank the owners, masters and crews of the ships that tow CPRs on a voluntary basis. Our appreciation goes to the anonymous reviewers for their constructive comments on our paper.

LITERATURE CITED

Batten SD, Allen RJS, Wotton COM (1998) The effects of the Sea Empress oil spill on the plankton of the southern Irish Sea. Mar Pollut Bull 36(10):764–774

Beaugrand G (2004a) Monitoring marine plankton ecosystems. I. Description of an ecosystem approach based on plankton indicators. Mar Ecol Prog Ser 269:69–81

Beaugrand G (2004b) Continuous Plankton Records: plankton atlas of the North Atlantic Ocean (1958–1999). I. Introduction and methodology. Mar Ecol Prog Ser Suppl 2004, CPR:3–10 (also available at www.int-res.com/journals/maps/cpr-plankton-atlas-2004/)

Beaugrand G, Edwards M (2001) Differences in performance among four indices used to evaluate diversity in planktonic ecosystems. Oceanol Acta 24:467–477

Beaugrand G, Ibañez F (2002) Spatial dependence of calanoid copepod diversity in the North Atlantic Ocean. Mar Ecol Prog Ser 232:197–211

Beaugrand G, Ibañez F (2004) Monitoring marine plankton ecosystems. II. Long-term changes in North Sea calanoid copepods in relation to hydro-climatic variability. Mar Ecol Prog Ser 284:35–47

Beaugrand G, Reid PC (2003) Long-term changes in phytoplankton, zooplankton and salmon related to climate. Global Change Biol 9(6):801–817

Beaugrand G, Reid PC, Ibañez F, Planque B (2000) Biodiversity of North Atlantic and North Sea calanoid copepods. Mar Ecol Prog Ser 204:299–303

Beaugrand G, Ibañez F, Lindley JA (2001) Geographical distribution and seasonal and diel changes in the diversity of calanoid copepods in the North Atlantic and North Sea. Mar Ecol Prog Ser 219:189–203

Beaugrand G, Reid PC, Ibañez F, Lindley JA, Edwards M (2002) Reorganization of North Atlantic marine copepod biodiversity and climate. Science 296:1692–1694

Beaugrand G, Brander KM, Lindley JA, Souissi S, Reid PC (2003) Plankton effect on cod recruitment in the North Sea. Nature 426:661–664

Brander KM, Dickson RR, Edwards M (2003) Use of Continuous Plankton Recorder information in support of marine management: applications in fisheries, environmental protection, and in the study of ecosystem response to environmental change. Prog Oceanogr 58(2-4):175–191

Centre for Environment Fisheries and Aquaculture Science (2000) Quality Status Report of the marine and coastal areas of the Irish Sea and Bristol Channel. Department of the Environment, Transport and the Regions, London

Colebrook JM (1964) Continuous Plankton Records: a principal components analysis of the geographical distribution of zooplankton. Bull Mar Ecol 6:78–100

Colebrook JM, Robinson GA (1961) The seasonal cycle of the plankton in the North Sea and northeastern Atlantic. J Cons Perm Int Explor Mer 26:156–165

CPR (Continuous Plankton Recorder) Survey Team (2004) Continuous Plankton Records: Plankton Atlas of the North Atlantic Ocean (1958–1999). II. Biogeographical charts. Mar Ecol Prog Ser Suppl 2004, CPR:11–75 (also available at www.int-res.com/journals/maps/cpr-plankton-atlas-2004/)

DeYoung B, Heath MR, Werner F, Chai F, Megrey B, Monfay P (2004) Challenges of modeling of ocean basin ecosystems. Science 304:1463–1466

Edwards M, John AWG (1998) Plankton. In: Barne JH, Robson CF, Kaznowska SS, et al. (eds) Coasts and seas of the United Kingdom, Region 7. South-east England: Lowestoft to Dungeness. Joint Nature Conservation Committee, Peterborough, p 81–83

Edwards M, Richardson AJ (2004) Impact of climate change on marine pelagic phenology and trophic mismatch. Nature 430(7002):881–884

Edwards M, John AWG, Johns DG, Reid PC (2001a) Case-history and persistence of the non-indigenous diatom

Coscinodiscus wailesii in the North-East Atlantic. J Mar Biol Assoc UK 81(2):207–211

Edwards M, Reid PC, Planque B (2001b) Long-term and regional variability of phytoplankton biomass in the Northeast Atlantic (1960–1995). ICES J Mar Sci 58(1):39–49

Edwards M, Richardson AJ, Batten S, John AWG (2004) Ecological Status Report: results from the CPR survey 2002/2003. SAHFOS Tech Rep 1:1–8 (ISSN 1744-0750)

Kirby RR, Reid PC (2001) PCR from the CPR offers historical perspective on marine population ecology. J Mar Biol Assoc UK 81(3):539–540

Lindley JA (1982) Continuous Plankton Records: geographical variations in numerical abundance, biomass and production of euphausiids in the North Atlantic Ocean and the North Sea. Mar Biol 71:7–10

NSTF (North Sea Task Force) (1993) North Sea Quality Status Report 1993: Oslo and Paris Commissions, London. Olsen & Olsen, Fredensbourg, Denmark

Reid PC (1999) The North Sea ecosystem: status report. In: Kumpf H, Steidinger K, Sherman K (eds) The Gulf of Mexico large marine ecosystem. Blackwell Science, Oxford, p 476–489

Reid PC, Battle EJV, Batten SD, Brander KM (2000) Impacts of fisheries on plankton community structure. ICES J Mar Sci 57(3):495–502

Reid PC, Colebrook JM, Matthews JBL, Aiken J (2003) The Continuous Plankton Recorder: concepts and history, from plankton indicator to undulating recorders. Prog Oceanogr 58(2-4):117–173

Reid PC, Edwards M, Beaugrand G, Stevens D, Wootton M (2004) State of the seas report: plankton. Contract report, UK Department of Environment, Food and Rural Affairs, London

Richardson AJ, Schoeman DS (2004) Climate impact on plankton ecosystems in the Northeast Atlantic. Science 305(5690):1609–1612

Richardson AJ, Walne A, John AWG, Jonas T, Lindley JA, Simms DW, Stevens D, Witt M (2006) Using Continous Plankton Recorder Data. Prog Oceanogr 68:27–74

Southward AJ and 15 others (2004) Long-term oceanographic and ecological research in the western English Channel. Adv Mar Biol 47:1–104

Stevens D, Reid PC (2004) History of the Continuous Plankton Recorder database. Workshop Report, Intergovernmental Oceanographic Commission, Paris, p 125–131

Thompson RC, Olsen Y, Mitchell RP, Davies A, Rowland SJ, John AWG, McGonigle D, Russell AE (2004) Lost at sea: Where is all the plastic? Science 304(5672):838

Vanden Berghe E, Brown M, Costello MJ, Heip C, Levitus S, Pissierssens P (eds) (2004) Proc 'The Colour of Ocean Data' Symp, Brussels, 25–27 November, 2002. IOC Workshop Report 188, UNESCO, Paris [and VLIZ Spec Publ 16]

Vezzulli L, Reid PC (2003) The CPR survey (1948–1997): a gridded database browser of plankton abundance in the North Sea. Prog Oceanogr 58(2-4):327–336

Vezzulli L, Dowland P, Reid PC, Clarke N, Papadaki M (2004) Gridded database browser of North Sea plankton: fifty years (1948–1997) of monthly plankton abundance from the Continuous Plankton Recorder (CPR) survey [CD-ROM]. Sir Alister Hardy Foundation for Ocean Science, Plymouth

Warner AJ, Hays GC (1994) Sampling by the Continuous Plankton Recorder survey. Prog Oceanogr 34(2-3):237–256

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: December 1, 2004; Accepted: December 12, 2005
Proofs received from author(s): June 19, 2006

Vol. 316: 257–268, 2006 | MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser | Published July 3

OPEN ACCESS

European marine biodiversity inventory and taxonomic resources: state of the art and gaps in knowledge

Mark J. Costello[1, 2,*], Philippe Bouchet[3], Chris S. Emblow[2], Anastasios Legakis[4]

[1] Leigh Marine Laboratory, University of Auckland, PO Box 349, Warkworth, New Zealand
[2] Ecological Consultancy Services Ltd (EcoServe), Unit B19 KCR Industrial Estate, Kimmage, Dublin 12, Ireland
[3] Muséum National d'Histoire Naturelle, Taxonomy-Collections Unit, CP51, 55 rue Buffon, 75005 Paris, France
[4] Zoological Museum, Department of Biology, University of Athens, Panepistimioupolis, Athens 15784, Greece

ABSTRACT: The European Register of Marine Species (ERMS) project has compiled a list of marine species in Europe and a bibliography of marine species identification guides. ERMS has also surveyed species identification and taxonomic expertise, and the state of marine species collections in Europe. A total of 29713 species-level taxa were catalogued from European seas. Overall, 90% of the taxon checklists were satisfactory, but non-halacarid Acarina, diatoms, lichens and cyanobacteria were not included, and geographical coverage of the European seas was incomplete for Rotifera and Brachiopoda. Lists that would benefit from further input include (1) those that have not yet been checked by an expert on European fauna, namely lists of the non-epicarid Isopoda, Cephalochordata, Appendicularia, Hemichordata, Hirudinea, Gnathostomulida, Ctenophora and Placozoa; (2) preliminary lists, including some of the above and lists of protists; and (3) lists with many species but which have been reviewed by only a few experts. These gaps are now being addressed in an online version of ERMS (www.marbef.org/data/erms.php). The bibliography of 842 identification guides shows that there are fewer guides for southern European seas, although they contain more species, than for those in northern Europe. Adequate guides for all of Europe's seas exist only for fishes. New guides are especially needed for the species-rich, but small-sized taxa, such as polychaete, oligochaete and turbellarian worms, and harpacticoid copepods. A database of >600 experts (individuals who stated themselves to be experts) and a subset of these recognised by their peers as being taxonomic experts was established. While there were generally more experts for taxa with a large number of species, there was no correlation between the number of taxonomists and the number of species per taxon; some taxa with thousands of species are studied by relatively few taxonomists. Such gaps in marine biodiversity knowledge and resources must be addressed by funding the production of additional species identification guides.

KEY WORDS: Database · Species · Taxonomy · Identification

INTRODUCTION

The unprecedented rate at which human activities around the world are causing species extinction is alarming (World Conservation Monitoring Centre 1992, Kirchner & Weil 2000). The patterns in the extinction of large predatory species on land are now occurring in oceans (Carlton 1993, Malakoff 1997, Casey & Myers 1998, Carlton et al. 1999, Roberts & Hawkins 1999, Baum et al. 2003, Myers & Worm 2003, 2005, Baum & Myers 2004). In addition, marine habitat degradation is reducing available living space and could lead to the extinction of other species. That extinctions are occurring before even half of the world's species have been described (May 1992, Barnes 1998, Gordon 2001) or named is evidence of a global information crisis. Such a gap in our knowledge of the world's biodiversity is thus a critical weakness in the world's 'knowledge

[*] Email: m.costello@auckland.ac.nz

economy'. The global economy is directly (e.g. food, materials) and indirectly (ecosystem services) dependent on biodiversity (Costanza et al. 1997, Costello 2000a, 2001). Considering that fisheries have never been so heavily harvested and that aquaculture is rapidly growing, one may expect concomitant growth in understanding what biodiversity exists in the oceans. However, there is no evidence of increased resources to identify and inventory marine biodiversity. Indeed, members of the scientific community have asserted that expertise in the form of taxonomists able to identify, describe and classify species is declining (e.g. Boero 2001, Giangrande 2003). However, these assertions are anecdotal and unsubstantiated. Where data have been provided, such as for Chile (Simonetti 1997), South Africa (Gibbons et al. 1999), the USA (Winston 1988) and globally (Diversitas 2000), they show that more taxonomists are needed to address the mismatch between the number of species in certain taxa and the corresponding number of taxonomists, but do not demonstrate a decline in the number of taxonomists. On the contrary, in Latin America, notably in Brazil, there has been increased employment and training in taxonomy since the 1980s (Carvalho et al. 2005).

Clearly, there have been insufficient taxonomic resources to describe the earth's present species, but there is no quantitative evidence of a decline in these resources in the available literature. Indeed, publications on marine biodiversity have been increasing in recent decades (Moustakas & Karakassis 2005). Because a taxonomist's reputation will grow during his/her lifetime, the fact that some taxonomic experts may be retired or nearing retirement may be more of a reflection of the fact that they have made lifetime contributions to the science than of the absence of successors; younger scientists have not had the time to build up a widely recognised reputation and thus may be overlooked when considering taxonomic expertise. Winston (1988) cites a study of taxonomists in the USA, which determined an average age of 44 yr, but felt it was biased by including students. Her 'impression' was that the average age was closer to 54 yr.

In addition to the age and number of taxonomic experts, other measures of taxonomic resources include the availability of organised species inventories, the currency of species identification guides and the condition of specimen collections. Fundamental to the management of any resource is an inventory of its parts and their abundance. The inventory of all species occurring in European seas (Costello et al. 2001) is the largest all-taxon marine species inventory available. This 'European Register of Marine Species' (ERMS) provides a means to indicate in which taxa most new species remain to be discovered; these findings may also be applicable globally. In the present paper, we compare the number of species per taxon in ERMS (Costello et al. 2001) to the available expertise.

The most basic requirement for people studying and working on aspects of biodiversity is the availability of species identification guides. Without such guides it is impractical for most people to know or study a group of species and consequently the biology, ecology and potential economic value of these species will remain unknown. People need rapid access to species identification guides, and funding agencies and publishers must know which guides are most urgently needed to fill taxonomic and geographic gaps. Thus, we analysed the taxonomic and geographic coverage of identification guides for marine species in Europe from a checklist we had previously compiled (Bouchet & Marmayou 2001).

The gathering of data in a standardised format facilitates gap analysis (e.g. Kelly & Costello 1996, Moustakas & Karakassis in press). Thus, in association with producing ERMS, we compiled (1) a database of expertise (including expert's age), (2) a catalogue of marine species identification guides (Bouchet & Marmayou 2001) and (3) a survey of museum collections (Legakis & Emblow 2003). Results of the survey of marine species collections have previously been reported (Legakis & Emblow 2003), but key points are also discussed in the present paper, to provide a more comprehensive review of the state of marine taxonomic resources in Europe.

MATERIALS AND METHODS

ERMS was a 2 yr project involving 22 organisations and 170 scientists (Costello 2000b). Groups of scientists within the project addressed the work described below, and others focused on communication with the scientific community and related organisations, including potential end-users.

Species lists. The ERMS project included species occurring from the strandline and 'splash zone' of the intertidal (littoral) through the subtidal (sublittoral) to the deep sea, including brackish waters to 0.5 salinity. The northern parts of the Baltic Sea are more freshwater than brackish, and it was left to the discretion of list compilers whether to include these species. The study area defined broadly as 'European seas' followed the database of European Mollusca (CLEMAM) (Fig. 1), and thus ranged from the North Pole along the east coast of Greenland to Iceland, along the mid-Atlantic ridge, across the 26° parallel to the coast of Africa, and into the Mediterranean and Black Seas. Inclusion of the islands of Madeira, Azores and Canaries brought sub-tropical species into ERMS; these had generally been excluded from previous reviews of European marine fauna and flora. Only taxonomically named species, and species whose occurrence

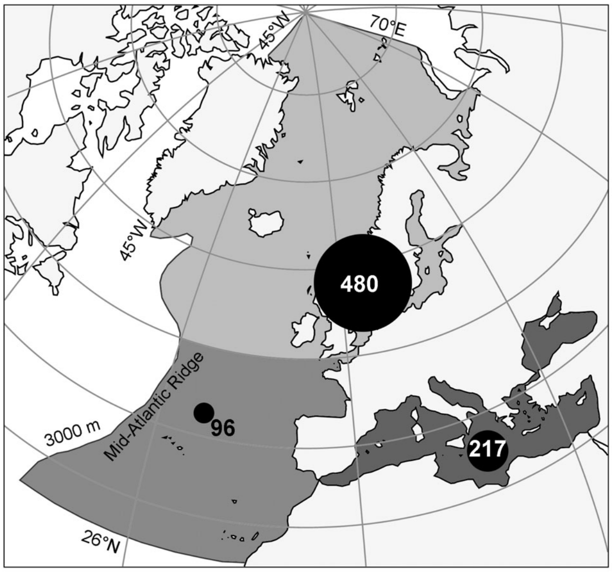


Fig. 1. Geographic scope of the European Register of Marine Species (ERMS) project. Numbers of identification guides for each of the 3 geographical areas are shown. In addition, 10 identification guides deal specifically with the NE Atlantic deep sea area

in the ERMS area had been previously published, were included. Synonyms and other names for a species were included in some instances. Saltmarsh angiosperm plants were excluded, as these are generally included in terrestrial plant inventories. Bacteria (Eubacteria and Archaea) were also excluded from the project. Where recognised, comments on the weaknesses of lists are made in the preface to that list (Costello et al. 2001). These comments, and criteria developed by M. J. Costello, were used to score the status of each list. The criteria were based on the source of information, expertise of the list compiler and involvement of >1 expertin compiling the list (Table 1).

Identification guides. Our review included identification guides with illustrations and keys to the larger metazoan groups and excluded (1) specialised literature dealing with a single genus or family; (2) non-illustrated checklists; (3) old literature that may be essential to a specialist, but is unobtainable to a general marine biologist; and (4) protists and microbia. It compiled all marine titles in major series (see Table 4), even if outdated or hard to obtain. Popular and semi-popular guides were listed separately, but they have not been comprehensively covered or included in the present analysis.

Table 1. Criteria used to indicate the quality of the species lists compiled during the ERMS (European Register of Marine Species) project, and the numbers of lists falling into these categories. For further information on scoring see 'Materials and methods'

Score	Criteria	No. of lists
+	Preliminary list, known to be or likely to be incomplete	4
++	Compiled from recent authoritative literature	7
+++	Compiled by expert in the group	28
++++	Checked by additional expert in the group	43
+++++	Checked by several experts in the group	31
	Total	113

Expertise. ERMS project participants supplied contact details for marine biologists from either their geographic or their taxonomic area. Lists of marine biologists from Britain, Germany, Greece, Italy, Ireland, Scandinavia, Spain and the western and eastern Mediterranean, as well as a list of European experts on algae, were received. In addition, lists of contact details for other marine biologists were obtained from the Internet. The focus of the project was on persons with expertise in marine species from countries of the European Union and the European Economic Area, so expertise in Eastern Europe has not been assessed. An initial list of 1200 people from 38 countries (29 European) with expertise in European marine species was compiled by the project. These people passed the list and an accompanying questionnaire on to an additional 160 colleagues who replied. Of the total of 614 respondents, 590 gave permission for their name to be entered in the database (i.e. they were still active and available for such work).

Each person in the database was asked to verify their contact details and provide their year of birth, the taxonomic groups in which they had expertise, their level of expertise, the geographic coverage of their expertise and their professional status. Requests were sent in the form of a standard questionnaire, with a summary of the project. A web-based submission form was also put on the web, and a general call for submissions made to various email discussion groups.

Because it proved difficult to set universal criteria to define a taxonomist, 2 registers were established: (1) persons with self-declared expertise in the identification of marine species in Europe and (2) peer-selected specialist or taxonomic experts in certain species groups. The persons producing lists of species for the project identified the latter 'taxonomic experts'.

RESULTS

Species lists

A total of 29 713 species-level taxa were catalogued from European seas, with the quality of information differing for different taxa. It was not expected that all

Table 2. Species lists in ERMS, the persons who compiled them and assisted in their compilation, the number of species per group and an indicator of how complete a list is of the described species (from Costello et al. 2001). C = confident of reasonable coverage of all European seas, including the Arctic, deep sea and Black Sea. See Table 1 for the status scoring system

Species group	Compiler (assisted by)	No. of species	Status	
Crytophytes	S. Brandt	14	+	
Euglenids Heterotrophic euglenoids	S. Brandt	26	+	
Haptophytes	S. Brandt	36	+	
Prasinophytes	S. Brandt	24	+	
Apicomplexa (free-living species)	S. Brandt	3	++	
Dinoflagellates	S. Brandt (M. Elbrächter)	718	++	
Kathablepharids	S. Brandt	2	++	
Placozoa	J. van der Land	2	++	
Ctenophora	J. van der Land	38	++	
Rotifera	M. O'Reilly	139	++	
Hirudinea	J. van der Land	36	++	
Thermosbaenacea	J. van der Land	2	++	
Isopoda excluding Epicaridea	J. van der Land	605	++	
Brachiopoda	C. Howson	18	++	
Appendicularia	J. van der Land	53	++	
Cephalochordata	J. van der Land	2	++	
Ciliates				
Aloricate oligotrichs	S. Agatha	82	+++	
Chonotricha	A. W. Jankowski	37	+++	
Folliculinids	M. Mulisch	30	+++	
Rhynchodida	A. W. Jankowski	42	+++	
Amoebae — testate	R. Meisterfeld	97	+++	
Apusomonads	S. Brandt	3	+++	
Choanoflagellates	S. Brandt	98	+++	
Euglenids — kinetoplastids	S. Brandt	13	+++	
Bicosoecids	S. Brandt	17	+++	
Labyrinthulids	M. Dick, S. Brandt	10	+++	
Thaustrochytrids	M. Dick, S. Brandt	15	+++	
Stramenopiles incertae sedis	S. Brandt	4	+++	
Thaumatomonads	S. Brandt	17	+++	
Protista incertae sedis (heterotrophic species)	S. Brandt	40	+++	
Mesozoa	J. Hallan, J. van der Land	36	+++	
Gnathostomulida	J. van der Land	25	+++	
Euphausiacea	J. van der Land	41	+++	
Hemichordata	J. van der Land	17	+++	
Fungi	N. Clipson, E. Landy, M. Otte (G. Bremer, G. Jones)	318	++++	
Amoebae — naked	A. Rogerson, A. Goodkov	74	++++	
Xenophyophora	O. Tendal, J. van der Land	20	++++	
Porifera	R. W. M. van Soest (N. Boury-Esnault)	1640	++++	
Siphonophora	G. M. Mapstone, J. van der Land (P. R. Pugh)	105	++++	
Chilopoda	A. Minelli	6	++++	
Diplopoda	A. Minelli	2	++++	
Insecta	A. Legakis	19	++++	
Phoronida	C. Emig	9	+++++	
Echiura	J. van der Land (J. I. Saiz-Salinas)	19	+++	C
Sipuncula	J. van der Land (J. I. Saiz-Salinas)	44	+++	C
Pentastomida	J. van der Land	2	+++	C
Stomatopoda	J. van der Land (P. Noel)	22	+++	C
Foraminifera	O. Gross	1167	++++	C
Actiniaria	J. H. den Hartog, J. van der Land (J. Ryland)	243	++++	C
Antipatharia	D. M. Opresko, J. van der Land	28	++++	C
Hydrozoa	W. Vervoort, S. D. Cairns, J. van der Land, P. Schuchert	684	++++	C
Gastrotrichia	J. L. D'Hondt, J. Van der Land	240	++++	C
Cephalorhyncha (=Loricifera, Priapulida, Kinorhyncha, Nematomorpha)	J. van der Land, B. Neuhaus	52	++++	C
Nematoda				
Free-living	G. De Smet, M. Vincx, A. Vanreusel, S. Vanhove, J. Vanaverbeke, M. Steyaert (F. Riemann)	1625	++++	C
Parasitic	D. Gibson (F. Moravec, H.-P. Fagerholm)	212	++++	C
Polychaeta	G. Bellan (C. Arvanitidis, J.-C. Dauvin, F. Gentil, G. Bachelet, H. Hansson, R. Barnick, D. Fiege, M. E. Petersen, T. Brattegard, T. Holthe)	1848	++++	C
Tardigrada	J. van der Land	76	++++	C
Pycnogonida	F. Krapp, J. Van der Land (J. Stock, R. Bamber, C. A. Child)	146	++++	C
Remipedia	G. Boxshall	1	++++	C
Branchiura	G. Boxshall	2	++++	C
Cladocera — Branchiopoda	G. Boxshall	9	++++	C
Mystacocarida	G. Boxshall	2	++++	C
Copepoda				
Calanoida	G. Boxshall	649	++++	C
Cyclopoida	G. Boxshall	177	++++	C
Harpacticoida	R. Huys	1357	++++	C
Misophrioida	G. Boxshall	16	++++	C
Monstrilloida	G. Boxshall	33	++++	C
Mormonilloida	G. Boxshall	2	++++	C
Platycopioida	G. Boxshall	3	++++	C
Poecilostomatoida	G. Boxshall (M. O'Reilly, D. Zavodnik)	353	++++	C
Siphonostomatoida	G. Boxshall	354	++++	C
Tantulocarida	G. Boxshall	13	++++	C
Cirripedia				
Non-parasitic Thoracica	A. Southward	107	++++	C
Parasitic Ascothoracida	G. Boxshall	10	++++	C
Parasitic Rhizocephala	G. Boxshall	28	++++	C

Table 2 (continued)

Species group	Compiler (assisted by)	No. of species	Status
Decapoda	M. Türkay	672	++++ C
Mysidacea	J. van der Land, T. Brattegard	198	++++ C
Isopoda, Epicaridea, Bopyridae	J. C. Markham	54	++++ C
Insecta			
Chironomidae	D. Murray	15	++++ C
Chaetognatha	H. Kapp, J. Van der Land	42	++++ C
Thaliacea	J. van der Land, R. Van Soest	35	++++ C
Macroalgae of Rhodophycota, Phaeophycota, Chlorophycota, and 2 genera of Xanthophycota	M. D. Guiry (G. Furnari, F. Rindi, E. Nic Dhonncha, S. Lawson)	1702	+++++ C
Seagrass	M. D. Guiry	5	+++++ C
Myxozoa	E. Karlsbakk	230	+++++ C
Octocorallia			
Pennatulacea	G. C. Williams, J. van der Land (K. Riemann-Zürneck)	37	+++++ C
Others	L. van Ofwegen, M. Grasshoff, J. van der Land	92	+++++ C
Scleractinia	S. D. Cairns, B. W. Hoeksema, J. van der Land (H. Zibrowius)	86	+++++ C
Cubozoa	P. Cornelius	1	+++++ C
Scyphozoa	P. Cornelius, G. Jarms, Y. M. Hirano, J. van der Land	53	+++++ C
Turbellaria	A. Faubel, C. Noreña	1137	+++++ C
Aspidogastrea	D. Gibson	4	+++++ C
Digenea	D. Gibson (M. Køoie, P. Bartoli)	592	+++++ C
Monogenea	R. Bray (L. Euzet, G. Kearn)	353	+++++ C
Cestoda	R. Bray (L. Euzet, B. B. Gorgiev)	312	+++++ C
Nemertea (Nemertini)	R. Gibson	478	+++++ C
Acanthocephala	D. Gibson (C. R. Kennedy, Z. M. Dimitrova)	67	+++++ C
Cycliophora	C. S. Emblow	1	+++++ C
Entoprocta	P. J. Hayward	45	+++++ C
Mollusca	S. Gofas, J. Le Renard, P. Bouchet, R. Giannuzzi-Savelli, A. Guerra, D. Heppell,	3353	+++++ C
Mollusca (continued)	T. Hoisaeter, E. Platts, S. Smith, J.-A. Sneli, A. Warén		
Oligochaeta	C. Erséus, B. Healy	190	+++++ C
Pogonophora	E. Southward, J. van der Land (T. Brattegard)	23	+++++ C
Acarina			
Halacaridae	I. Bartsch	214	+++++ C
Ostracoda	D. Horne, A. Bruce, J. Whittaker	769	+++++ C
Amphipoda	D. Bellan-Santini, M. J. Costello (S. Ruffo, J.-C. Dauvin, L. Collier)	1183	+++++ C
Cumacea	L. Watling (T. Brattegard)	188	+++++ C
Tanaidacea	G. Bird (M. Gutu)	280	+++++ C
Bryozoa	P. J. Hayward (J. Harmelin)	724	+++++ C
Echinodermata	H. G. Hansson (S. Stöhr, C. Massin, A. Gebruk, A. Mironov, A. Smirnov, D. Zavodnik, M. Garrido)	648	+++++ C
Ascidiacea & Sorberacea	C. Monniot, D. Connor, P. Lozouet	393	+++++ C
Pisces			
Agnatha	J. van der Land, M. J. Costello (L. Collier)	5	+++++ C
Chondrichthyes	J. van der Land, M. J. Costello, R. Serrão Santos and F. Mora Porteiro. (L. Collier)	145	+++++ C
Osteichthyes	J. van der Land, M. J. Costello, R. Serrão Santos F. Mora Porteiro (L. Collier)	1199	+++++ C
Tetrapoda			
Aves	J. van der Land, M. Ramos, J. Templado	74	+++++ C
Reptilia	J. van der Land, M. Ramos, J. Templado	5	+++++ C
Mammalia	J. van der Land, M. Ramos, J. Templado	50	+++++ C

lists could be produced to the same standard, because of the varying availability of recently published reviews and of expertise. Only 4% of the taxonomic lists are considered incomplete, representing probably ≤2% of the total number of described species (Tables 1 & 2). Lists with scores >2 (indicated by a corresponding number of plus signs) were considered satisfactory, and 90% of all lists were in this category. However, 63% of the lists (scores of 3 and 4) would benefit from further expert review. Non-halacarid Acarina, diatoms, lichens and cyanobacteria were not compiled, and geographical coverage of the European seas was incomplete for Rotifera and Brachiopoda. Lists that were satisfactory, but that would benefit from further input include (1) lists that had not been checked by an expert on European fauna, namely lists for the non-epicarid Isopoda, Cephalochordata, Appendicularia, Hemichordata, Hirudinea, Gnathostomulida and Ctenophora and (2) lists known to be preliminary, including some of the above and several for protists.

Lists with many species merit further attention because it is very likely that these groups will contain species newly described to science, and/or changes in nomenclature, within a short time. The lists of macroalgae, Porifera and Mollusca were derived from well-established databases, and the lists of fishes were cross-checked against other world-wide listings. However, other large lists were prepared for the first time for this project. Because of the size of these lists, no single person can be an expert on all of the species covered, and the editorial task per person is greater. Thus, the lists of Polychaeta, Amphipoda, Harpacticoida and Turbellaria may benefit from further review.

Identification guides

Of the 842 identification guides compiled, 362 titles (43%) have been published in national or regional series, some dealing specifically with marine fauna and flora (Table 3). Although volumes may be obsolete or hard to obtain (Table 4), these series are often the guides most frequently used by non-specialists attempting to identify marine species in Europe. The 'Synopses of the British Fauna' was the most comprehensive series; it was estimated that it covered 80% of the species encountered in northern and Arctic waters, and 50% of the species encountered in the Mediterranean and the Atlantic archipelagos. One series was limited to the seaweeds of the British Isles. For the Mediterranean, the most complete coverage was by Faune de France and Fauna e Flora del Golfo di Napoli, but these series are now largely obsolete. While Fauna Iberica has a number of titles in preparation, the eastern Mediterranean remains poorly covered. No series has comprehensively covered the major groups of macrobenthos from the Arctic to the Mediterranean, so accurate identification of these taxa relies on a patchwork of guides of uneven reliability and relevance to the area concerned.

The geographical coverage was very uneven (Table 4), with 52% of the titles particular to northern Europe (the British Isles, North Sea and Scandinavia), 22% to the Mediterranean and 11% to the Atlantic–Lusitanian region (Bay of Biscay to Morocco and the Atlantic archipelagos). (The total does not add up to 100% because some general guides have not been allocated to a geographical region.) No series considered the deep-sea fauna in particular. Guides to deep-sea fauna were lim-

Table 3. Adequacy of identification guides in northern European, other Atlantic (including Lusitanian) waters, and Mediterranean and Black Seas, compared with the number of species recorded by ERMS for each taxon **: recent; *: out of date but useful; –: no useful guides; shaded areas indicate where guides do not exist

Taxon		Northern Europe	Atlantic	Medit. and Black Sea	No. of species
Acanthocephala		–	–	–	67
Annelida	Hirudinea	–	–	–	*36*
	Oligochaeta	–	–	*	*190*
	Polychaeta	**	–	–	1848
Brachiopoda		**	–	**	*>18*
Bryozoa		**	–	**	724
Cephalochordata		*	–	–	2
Chaetognatha		**	–	–	42
Chelicerata	Halacarida	**	–	–	>214
	Pycnogonida	**	–	–	146
Cnidaria		**	**	**	1224
Crustacea	Branchiopoda	–	–	–	*9*
	Cirripedia	**	–	**	*145*
	Copepoda	**	–	–	*2957*
	Ostracoda	**	–	–	*769*
	Stomatopoda	**	–	**	*22*
	Mysidacea	*	–	–	*198*
	Amphipoda	**	–	**	*1183*
	Isopoda	**	–	–	*659*
	Tanaidacea	**	–	–	*280*
	Cumacea	**	–	–	*188*
	Decapoda	**	**	*	*672*
Ctenophora		–	–	–	38
Echinodermata		**	**	–	648
Entoprocta		**	–	–	45
Foraminifera		–	–	–	1167
Gastrotricha		–	–	–	240
Hemichordata		–	–	–	17
Insecta		–	–	–	34
Kinorhyncha		–	–	–	41
Mollusca		**	**	**	3353
Nematoda		**	–	–	1837
Nematomorpha		–	–	–	3
Nemertea		**	–	–	478
Phoronida		**	–	–	9
Platyhelminthes		**	–	–	2398
Pogonophora		–	–	–	23
Porifera		**	**	–	1640
Rotifera		**	–	**	>139
Sipunculida		**	**	**	44
Tardigrada		–	–	–	76
Tunicata		**	*	**	393
Vertebrata	Pisces	**	**	**	*1349*
	Reptilia	**	**	**	*5*
	Aves	**	**	**	*74*
	Mammalia	**	**	**	*50*
Flora		**	**	**	1707

Table 4. Summary of the titles, number of issues concerning marine species, currency (years published), language and coverage of marine species in the respective area, for the major series of identification guides that include some marine species in Europe

Series title	Area	Marine issues	Years	Language	Marine (%)
Danmarks Fauna	North Sea, Baltic Sea	17	1910–1996	Danish	40
Die Tierwelt Deutschlands	North Sea, Baltic Sea	29	1925–1996	German	60
Fauna d'Italia	Central and western Mediterranean	6	1956–1986	Italian	5
Fauna e Flora del Golfo di Napoli	Western Mediterranean	39	1880–1982	Italian	60
Fauna Graeciae	Eastern Mediterranean	3	1988–1996	English	2
Fauna Iberica	Atlantic France, Iberia and western Mediterranean	3	1992–1996	Spanish	2
Fauna Marinha de Portugal	Atlantic France and Iberia	10	1931–1936	Portuguese	<1
Fauna Republicii Socialiste România (or Fauna Republicii Populare Romîne)	Black Sea	14	1941–1983	Romanian	20
Fauna SSSR/Opredeliteli po Faune SSSR	Baltic Sea, White Sea and adjacent Arctic waters	55	1932–1996	Russian	40
Fauna van Nederland	North Sea	9	1932–1956	Dutch	20
Faune de France	NE Atlantic to Norway and western Mediterranean	16	1923–1966	French	20
Guide per il riconoscimento delle specie animale delle acque lagunari e costiere italiane	Central and western Mediterranean	11	1980–1983	Italian	15
Marine Invertebrates of Scandinavia	North Sea, Baltic Sea, Arctic	10	1966–1998	English	5
Seaweeds of the British Isles	NE Atlantic	7	1977–1994	English	70
Synopses of the British Fauna	NE Atlantic	44	1944–1998	English	60
Tierwelt der Nord- und Ostsee	North Sea, Baltic Sea	71	1925–1958	German	80

ited to selected groups of Mollusca (4), Tunicata (3), Crustacea (2) and Echinodermata (1).

Taxonomic coverage was equally uneven, with current and comprehensive identification guides available only for the vertebrates of all of Europe's seas (Table 3). Other taxa that were well covered in recent guides were Mollusca, Cnidaria and Sipunculida. New guides are especially needed for the species-rich, but small-sized taxa, such as: (1) the worms Polychaeta, Oligochaeta, Nematoda, Nemerta and Platyhelminthes (Turbellaria, parasitic Digenea and Monogenea); (2) the crustaceans Copepoda, Ostracoda, Isopoda, Tanaidacea and Cumacea; and (3) the Foraminifera.

The number of guides published annually increased from the 1950s to 1980s in line with general publication trends (Fig. 2).

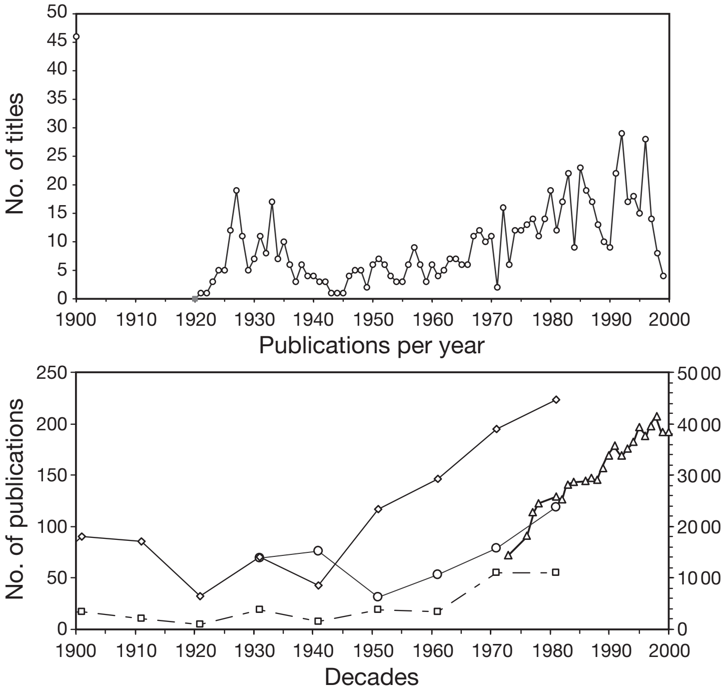


Fig. 2. (a) Total number of identification guides to European marine fauna and flora published per year and (b) number of publications in Irish periodicals (diamonds, Kelly & Costello 1996), identification guides in Europe (circles, present study), on amphipod crustaceans in Ireland (squares, Costello et al. 1990) and in ASFA (Aquatic Science and Fisheries Abstracts) (triangles, righthand *y*-axis, Moustakas & Karakassis 2005)

Expertise

The level of response to the survey of expertise of 37 % is considered very good, because a significant number of the persons contacted may no longer have been at the

Table 5. Country and number of respondents (n) in each country, including countries not completely covered by initial lists as they were outside the study area (*)

Country	n	Country	n	Country	n
Australia*	3	Ireland	13	Seychelles*	1
Austria	5	Israel*	7	Slovenia*	1
Belgium	13	Italy	51	South Africa*	1
Brazil*	1	Japan*	1	Spain	95
Bulgaria*	2	Lebanon*	2	Sweden	11
Croatia*	7	Malta*	1	Switzerland	1
Denmark	18	New Zealand*	1	The Netherlands	12
Egypt*	7	Northern Ireland	3	Turkey*	8
Finland	5	Norway	33	UK	65
France	37	Poland	12	Ukraine*	2
Germany	62	Portugal	11	USA*	14
Greece	36	Romania*	3	Venezuela*	1
Iceland	4	Russia*	40		

Table 6. Breakdown of people by employment status

Status	No. of respondents	% of respondents	Average age
Student	19	3	30.6
Non-professional	7	1	50.2
Professional (private sector)	33	6	41.8
Professional (public service/academic)	472	80	46.6
Retired professional	29	5	67.0

address used, due to job changes, retirement, or death. The response rate of 54% of people contacted via email was almost twice that of people contacted by post. However, some email contacts replied by fax or post. It is notable that 26% of respondents were not contacted directly by the project. This suggests that despite efforts of the project team to compile individual contact details, a number of experts may still be missing from the database.

The register contained people from 29 European and 9 non-European countries (Table 5). The number of respondents was higher from countries for which lists of marine biologists had previously been compiled. Other countries had fewer respondents, as did countries which were not initially targeted by the project, in particular the Eastern European and non-European countries. Countries with the best coverage of taxonomic groups were those which were sufficiently represented in previous lists of experts. The majority of respondents stated that they had global (245) or regional (375) expertise, whilst only 113 felt they were limited to local expertise.

The age structure of respondents showed young students were clearly distinguished from older, retired professionals (Fig. 3, Table 6). The youngest person was 23 and the eldest 89 yr old. The average age was 47 yr. Of the respondents, 80% were professionals in the public service or academic sector (Table 6).

Although there were >100 people with expertise in the identification of Arthropoda (largely Crustacea) and algae (Table 7), we do not know how many are able to identify the more taxonomically difficult taxa within these groups. In our distinction between identification and taxonomic expertise, we found that there was a positive relationship between the number of people with expertise in species identification and the number of species in the phyla (Fig. 4). In contrast, the number of taxonomic experts did not correlate as well with the number of species (Fig. 4). The number of taxonomic experts was generally lower than the number of identification experts for the phyla compared, with the exception of Porifera. The numbers for Bryozoa, Phoronida and Platyhelminthes were similar for both types of experts, suggesting that only taxonomic experts identified these taxa.

DISCUSSION

Species lists

The updating of ERMS is a continuous process as new discoveries are made and nomenclature changes. This requires a management structure that is sustainable in long-term rather than project-by-project

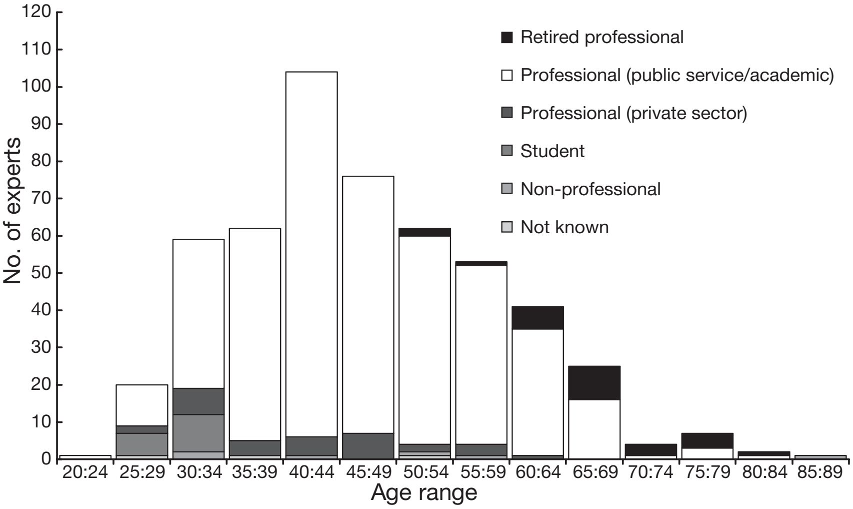


Fig. 3. Age distribution in 1999, and employment status, of people with expertise in marine species identification

Table 7. Number of identification experts by phyla and higher level taxonomic groups (e.g. algae are not a phylum) in the register. Some people have expertise in >1 phylum

Taxon	No. of experts	Taxon	No. of experts	Taxon	No. of experts
Acanthocephala	4	Cnidaria	22	Nematoda	15
Algae	121	Ctenophora	2	Nemertini	2
Annelida	63	Cyanophyta	11	Phoronida	4
Arthropoda	155	Cycliophora	2	Pisces	55
Bacteria	5	Echinodermata	24	Platyhelminthes	7
Brachiopoda	11	Echiurida	4	Porifera	14
Bryozoa	16	Entoprocta	1	Protista	24
Cephalorhyncha	10	Fungi	1	Rhodophyta	73
Chaetognatha	5	Gastrotricha	4	Rotatoria	1
Chlorophyta	72	Gnathostomulida	1	Sipuncula	4
Chordata	15	Granuloreticulosa	8	Spermatophyta	2
Chromophyta	82	Mesozoa	1	Tardigrada	2
Ciliata	5	Mollusca	82	Urochordata	11
				Vertebrata	13

management. Thus, a legal organisation was established by the ERMS project called the 'Society for the Management of European Biodiversity Data' (www.smebd.org) (Costello 2000b). All persons who make intellectual contributions to ERMS are life members, and they authorise the society to own and manage ERMS on behalf of the scientific community. Members elect a governing council that authorises where the top-copy of ERMS is hosted and appoints an editorial committee (the ERMS Executive Committee) to make the day-to-day decisions regarding administrative changes. The society may also facilitate the rescue of 'orphaned' biodiversity databases (e.g. where a scientist has retired and there is no successor to maintain the database) by finding suitable new hosts or managers for them. The society's ERMS Executive Committee has established an editorial board responsible for the quality control and development of ERMS. In this way, members of the board, including all taxonomic experts responsible for keeping taxonomic nomenclature within ERMS current, perform a role analogous to that of the editorial board of a scientific journal. Similarly, their time is contributed as part of their service to science, a view supported by the Consortium of European Taxonomic Facilities (2004). However, unlike paper publications, ERMS is dynamic in that errors can be corrected and new findings added, and species names can be directly connected to websites with more information about them. Thus, the Internet-accessible publication of ERMS will improve in quality and comprehensiveness over time. Where special costs arise, such as in converting ERMS to a relational database that can be edited online and making it interoperable with other databases, special project funding is sought, such as that provided by the Marine Biodiversity and Ecosystem Functioning (MarBEF project (www.marbef.org). Through MarBEF, ERMS 2.0 is being produced; this will be more complete taxonomically, will have more associated information and will also be freely available on the World Wide Web (Costello 2004). Thus, most of the gaps in ERMS that have been identified in the present paper (Table 2) will be addressed. These solutions to the long-term development and quality assurance of ERMS appear to be unparalleled in other online data resources, and merit consideration for other scientific endeavours.

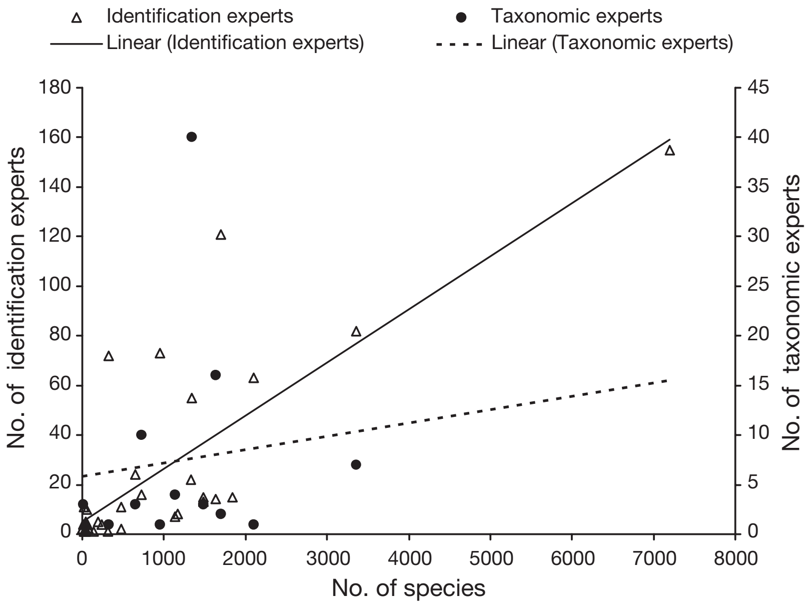


Fig. 4. Scatter plot of number of species against number of identification (triangles, solid line) and taxonomic (solid dots, dashed line) experts in the register

A global inventory of species exists for only about 20% of the estimated number of described species (Bisby et al. 2004), and we estimate that the proportion for marine species is not much greater. Without a complete inventory of a taxon it is difficult for people to know if the specimens they look at have already been described. Indeed, the time taxonomists spend describing species they were not aware had already been described, and/or correcting such mistakes, would be saved if taxonomic nomenclatures were more widely and rapidly available, and if new species were promptly registered on the Internet. If this were done, links could be made to publications, experts and other websites, and mistakes could be corrected quickly instead of waiting decades for another paper publication.

Guides

There are both taxonomic and geographic gaps in the availability of up-to-date marine species identification guides. The large, common, and/or ecologically significant species are covered in several to many guides. In contrast, many of the smaller, rarer or taxonomically difficult to identify species are not covered in any of the guides listed. Yet, these species may be of great importance to biodiversity, ecosystem function and marine resources. Although many identification guides are available for those regions of Europe in which the marine fauna and flora is least diverse (the North and Baltic Seas), there are considerably fewer guides for those regions of Europe in which the marine fauna and flora is most diverse (the Mediterranean, the Atlantic archipelagos, the deep sea). Thus, the taxonomic and geographic gaps that most urgently require attention are the smaller sized taxa in the southern European seas (both Atlantic and Mediterranean).

Although the number of guides published has increased since the 1940s, so has the number of scientific publications in general. From the 1930s to 1990s, our data showed an 88% increase in the number of identification guides. However, similar increases of 89 and 92% for marine publications on amphipod crustaceans (Costello et al. 1990) and biology periodicals in Ireland (Kelly & Costello 1996), respectively, and marine publications in 'Aquatic Sciences and Fisheries Abstracts' (Moustakas & Karakassis 2005) suggest that the number of identification guides has not increased relative to other marine biology publications (Fig. 2). One reason why fewer identification guides are being produced is that evaluations for careers and funding rely strongly on impact factor, which favours multiple publications in journals rather than as books or volumes in series.

Identification guides are not only important as a resource, but also because they enable many more scientists to begin to recognise and study previously little known species. Thus, plots of the accumulated rate of discovery of marine species (e.g. Costello et al. 1996) have a sigmoid shape, from little discovery in the early stages, to rapid discovery once some guides have been published, and then decreasing rates of discovery as the taxa become well known. Unfortunately, for European marine species as a whole, these discovery rates are still in the second stage, and the point at which all species will be described is still nowhere in sight (M. J. Costello unpubl. data).

Expertise

It is difficult to assess expertise whether by self-evaluation or by peer evaluation, because people can only assess based on what they know about the limitations of their own skills and the skills of others, and these views may differ across taxonomic groups (e.g. not many people describe new species of mammals and birds). An identification expert in taxa that are difficult to identify may have more skills than a taxonomic expert in a species group with a few easily identified species. Despite these problems, it was possible to assess the relative abundance of identification and taxonomic experts in European marine species. It was also possible to identify that gaps do exist in identification and taxonomic expertise with respect to European marine species, and that further work should be focused where gaps exist.

Our results, which gave an average age of 47 yr for experts in the identification of European marine species, supports Winston's (1988) view that an average age of 44 for taxonomists in a US survey was an underestimate due to inclusion of students. However, our survey probably under-sampled younger scientists (Fig. 3) because they are less well-known and more mobile, and does not support Winston's (1988) view that their average age may be closer to 54. Neither does it suggest that most taxonomists will be retiring within the next 10 yr.

For taxa that are considered to have a relatively complete list, the number of identification experts for each group was higher than the number of taxonomic experts. However, in some phyla (e.g. Porifera), classes and orders, more taxonomic experts than identification experts were listed. When the number of identification experts or taxonomists is similar, it suggests that only taxonomists work on these groups, and that they are not widely studied by ecologists (we assume that the majority of non-taxonomist species identification experts are ecologists). In the case of readily identifiable taxa it may be correct to assume that all taxonomic experts were also identification experts, while a number of identification experts would not be taxonomists. However, in the case of taxa that are difficult to identify without the use of specialised techniques (e.g. Porifera), this may not be the case, and the numbers of taxonomic experts are closer to the number of identification experts.

The age distribution did not indicate any imminent extinction of identification expertise. The peer-selected top experts in taxonomy are at later stages in their careers, because their publication record, expertise and peers' knowledge of their expertise will increase over time. Thus, it may always be the case that the leading taxonomists will be nearing retirement. Some of the identification experts will include younger people, who will be able to do taxonomic work as the need arises. However, while our data do not support the common assertion that there is a danger of losing taxonomic expertise, they do identify important gaps in expertise that must be filled by new positions if biodiversity is to be discovered, conserved and used sustainably. There were generally more people identifying taxa with more species, although there was no correlation between the number of taxonomists and species in their taxa. It was evident that some taxa with thousands of species have insufficient taxonomists. Thus, a mismatch between taxonomic need and expertise exists for European marine species, as has been found across all taxa in Chile (Simonetti 1997), South Africa (Gibbons et al. 1999), the USA (Winston 1988) and globally (Diversitas 2000). New species reported in the online Zoological Record in 2002 and 2003 include 118 from off the coasts of southern Europe (Mediterranean, Black Sea, Iberia, Canary Islands, Azores, Madeira), 88 from the Atlantic coasts of western Europe, and 36 from Arctic Europe. However, 36% of the Mediterranean species were described from Italy, and 25% of the Atlantic species from Spain, illustrating the relative strength of taxonomy in these countries. This may be an indication that the geographic mismatch between species richness and the need for taxonomic effort has begun to be addressed in Europe.

Collections

Both large and small collections of marine species shared a common problem—insufficient resources for proper maintenance (Legakis & Emblow 2003). Most (64%) of the collections were incompletely catalogued, and only 10% had their catalogue in electronic form (Table 8). New funding is therefore essential if the knowledge included in the collections is to be available on the Internet. Almost half the collections had specimens from around the world, and making information from collections available through the Internet would help share and repatriate this knowledge to the source countries.

Table 8. Presence and extent of coverage of collection catalogues in paper and electronic (computerised) form (data from Legakis & Emblow 2003)

	Paper		Electronic	
	No.	%	No.	%
Full coverage	29	36	8	10
Part coverage	35	44	38	54
No coverage	16	20	31	36

Sourcing type specimens through online databases will facilitate the production of guides and taxonomic training. Collection managers should include electronic databases as part of the routine management of their collections and seek special funding to help integrate past collection knowledge into such databases, as demonstrated by Martin et al. (2004).

CONCLUSIONS

A priority for further infrastructure research should be the production of guides for the identification of species, especially those taxa prioritised in the present study. Their preparation will require increased funding of taxonomic research into areas of European seas where most species have yet to be described. Funding may be direct, such as through the US National Science Foundation's PEET (Partnerships for Enhancing Expertise in Taxonomy) programme, or indirect, through ecological, fisheries, informatics and molecular research projects, including funds for the necessary supporting taxonomic research and infrastructure. This will have the 2-fold benefit of providing employment for taxonomists to produce the guides, and the guides will enable many others to be trained to identify and work with the species covered. These guides should not only illustrate and describe the species, but review existing knowledge on their habitat and distribution (e.g. as done by the Synopses of the British Fauna at present). This information should be available electronically and help extend the ERMS species register into a species information system. These guides could be published on the Internet, a compact disc, and/or as a book. The advantage of electronic publication is that species identification is possible through electronic keys that can be more user-friendly and functional than traditional paper keys. Thus, 'biodiversity informatics', the use of information technology in biodiversity data management, must join molecular techniques (e.g. DNA bar-coding) as a new tool in taxonomy. Informatics and molecular tools are not alternatives to taxonomy. The need to identify species in practical ways still requires taxonomic descriptions and images, type specimens as standards for comparative analysis, and a species naming system that enables communication of 'what it is'. Biodiversity informatics can increase the visibility and availability of taxonomic knowledge and its associated data, thereby facilitating more cost-effective use of resources.

Acknowledgements. The ERMS project was partially funded by the European Commission Marine Science and Technology (MAST) research programme (Project Number MAS3-CT97-0146). We thank the many participants in ERMS for their collaboration and support, and Camilo Mora, Gordon Patterson, Christos Arvantidis, Daphne Fautin, Yannis Karakassis and anonymous referees for helpful comments on this paper.

LITERATURE CITED

Barnes RD (1998) Diversity of organisms: How much do we know? Am Zool 29:1075–1084

Baum JK, Myers RA (2004) Shifting baselines and the decline of pelagic sharks in the Gulf of Mexico. Ecol Lett 7:135–145

Baum JK, Myers RA, Kohler DG, Worm B, Harley SJ, Doherty PA (2003) Collapse and conservation of shark populations in the northwest Atlantic. Science 299:389–392

Bisby FA, Froese R, Ruggiero MA, Wilson KL (2004) Species 2000 & ITIS catalogue of life: 2004 annual checklist. Indexing the world's known species, CD-ROM. Species 2000, Los Baños

Boero F (2001) Light after dark: the partnership for enhancing expertise in taxonomy. Trends Ecol Evol 16:266

Bouchet P, Marmayou J (2001) Bibliography of identification guides to marine species in Europe. In: Costello MJ, Emblow C, White R (eds) European Register of Marine Species. A check-list of marine species in Europe and a bibliography of guides to their identification. Patrimoines Naturels 50:377–396

Carlton JT (1993) Neoextinctions of marine invertebrates. Am Zool 33:499–509

Carlton JT, Geller JB, Reaka-Kudla ML, Norse EA (1999) Historical extinctions in the sea. Annu Rev Ecol Syst 30: 515–538

Carvalho MR de, Bockmann FA, Amorim DS, Vivo M de and 12 others (2005) Revisiting the taxonomic impediment. Science 307:353

Casey JM, Myers RA (1998) Near extinction of a large widely distributed fish. Science 281:690–692

Consortium of European Taxonomic Facilities (2004) CETAF expresses its support for European directories of species names. Available at www.cetaf.org/Spec.database.pdf

Costanza R, d'Arge R, de Groot R, Farber S and 9 others (1997) The value of the world's ecosystem services and natural capital. Nature 387:253–260

Costello MJ (2000a) A framework for an action plan on marine biodiversity in Ireland. Mar Res Ser (Mar Inst Ir) 14:1–47

Costello MJ (2000b) Developing species information systems: the European Register of Marine Species. Oceanography 13:48–55

Costello MJ (2001) To know, research, manage, and conserve marine biodiversity. Oceanis 24:25–49

Costello MJ (2004) A new infrastructure for marine biology in Europe: marine biodiversity informatics. MARBEF Newsl 1:22–24

Costello MJ, Holmes JMC, McGrath D, Myers AA (1990) A review and catalogue of amphipod Crustacea in Ireland. Ir Fish Invest B 33:1–70

Costello MJ, Emblow CS, Picton BE (1996) Long term trends in the discovery of marine species new to science which occur in Britain and Ireland. J Mar Biol Assoc UK 76: 255–257

Costello MJ, Emblow C, White R (eds) (2001) European Register of Marine Species. A check-list of marine species in Europe and a bibliography of guides to their identification. Patrimoines Naturels 50:1–463

Diversitas (2000) Implementing the GTI: recommendations from Diversitas core programme Element 3, including an assessment of present knowledge of key species groups. International Union of Biological Sciences, Paris

Giangrande A (2003) Biodiversity, conservation and the 'taxonomic impediment'. Aquat Conserv 13:451–459

Gibbons MJ, and 62 others (1999) The taxonomic richness of South Africa's marine fauna: a crisis at hand. S Afr J Sci 95:8–12

Gordon DP (2001) Marine biodiversity. Alpha 108:1–8

Kelly KS, Costello MJ (1996) Temporal trends and gaps in marine publications in Irish periodicals. In: Keegan BF O'Connor R (eds) Irish marine science 1995. Galway University Press, p 37–48

Kirchner JW, Weil A (2000) Delayed biological recovery from extinctions throughout the fossil record. Nature 404: 177–180

Legakis A, Emblow CS (2003) The register of collections of European marine species: an overview. In: Legakis A, Sfenthourakis S, Polymeni R, Thessalou-Legaki M (eds) The new panorama of animal evolution. Proc 18th international congress of zoology. Pensoft, Sofia, p 603–609

Malakoff D (1997) Extinction on the high seas. Science 277: 486–488

Martin A, Van Guelpen L, Pohle G, Costello MJ (2004) Development of an Atlantic Canada marine species information system based on a museum collection: a case study. In: Vanden Berghe E, Brown M, Costello MJ, Heip C, Pissierssens P (eds) Proc 'The Colour of Ocean Data' Symp Brussels 25–27 November 2002. IOC Workshop Report 188, UNESCO, Paris (and VLIZ Spec Publ 16), p 71–76

May RM (1992) Bottoms up for the oceans. Nature 357: 278–279

Moustakas A, Karakassis I (2005) How diverse is aquatic biodiversity research? Aquat Ecol 39:367–375

Myers RA, Worm B (2003) Rapid worldwide depletion of predatory fish communities. Nature 423:280–283

Myers RA, Worm B (2005) Extinction, survival, or recovery of large predatory fishes. Philos Trans R Soc Lond B 360: 13–20

Roberts CM, Hawkins JH (1999) Extinction risk in the sea. Trends Ecol Evol 14:241–246

Simonetti JA (1997) Biodiversity and a taxonomy of Chilean taxonomists. Biodiversity Conserv 6:633–637

Winston JE (1988) The systematist's perspective. Mem South Calif Acad Sci 13:1–6

World Conservation Monitoring Centre (1992) Global biodiversity: status of the Earth's living resources. Chapman and Hall, London

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: February 11, 2005; Accepted: September 9, 2005
Proofs received from author(s): June 2, 2006

Vol. 316: 269–283, 2006

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

Modeling habitat distribution from organism occurrences and environmental data: case study using anemonefishes and their sea anemone hosts

J. M. Guinotte[1,2,5,*], J. D. Bartley[1], A. Iqbal[1], D. G. Fautin[1,3], R. W. Buddemeier[1,4]

[1]Kansas Geological Survey, 1930 Constant Avenue, Lawrence, Kansas 66047, USA
[2]School of Tropical Environmental Studies and Geography, James Cook University, Townsville, Queensland 4810, Australia
[3]Department of Ecology and Evolutionary Biology, and Natural History Museum and Biodiversity Research Center, and
[4]Department of Geography, University of Kansas, Lawrence, Kansas 66045, USA

[5]*Present address:* Marine Conservation Biology Institute, 2122 112th Ave NE, Suite B-300, Bellevue, Washington 98004, USA

ABSTRACT: We demonstrate the KGSMapper (Kansas Geological Survey Mapper), a straightforward, web-based biogeographic tool that uses environmental conditions of places where members of a taxon are known to occur to find other places containing suitable habitat for them. Using occurrence data for anemonefishes or their host sea anemones, and data for environmental parameters, we generated maps of suitable habitat for the organisms. The fact that the fishes are obligate symbionts of the anemones allowed us to validate the KGSMapper output: we were able to compare the inferred occurrence of the organism to that of the actual occurrence of its symbiont. Characterizing suitable habitat for these organisms in the Indo-West Pacific, the region where they naturally occur, can be used to guide conservation efforts, field work, etc.; defining suitable habitat for them in the Atlantic and eastern Pacific is relevant to identifying areas vulnerable to biological invasions. We advocate distinguishing between these 2 sorts of model output, terming the former maps of realized habitat and the latter maps of potential habitat. Creation of a niche model requires adding biotic data to the environmental data used for habitat maps: we included data on fish occurrences to infer anemone distribution and vice versa. Altering the selection of environmental variables allowed us to investigate which variables may exert the most influence on organism distribution. Adding variables does not necessarily improve precision of the model output. KGSMapper output distinguishes areas that fall within 1 standard deviation (SD) of the mean environmental variable values for places where members of the taxon occur, within 2 SD, and within the entire range of values; eliminating outliers or data known to be imprecise or inaccurate improved output precision mainly in the 2 SD range and beyond. Thus, KGSMapper is robust in the face of questionable data, offering the user a way to recognize and clean such data. It also functions well with sparse datasets. These features make it useful for biogeographic meta-analyses with the diverse, distributed datasets that are typical for marine organisms lacking direct commercial value.

KEY WORDS: Biogeography · Clownfish · Ecological niche · Range · GIS

INTRODUCTION

Biogeographic information, whether about taxa, guilds, or groups of associated organisms, is fundamental to human use and understanding of the environment. Electronic resources are rapidly enhancing the volume and diversity of information that can be brought to bear on problems such as the identification and protection of biodiversity, actual or potential invasive species, and diagnosis and prediction of the effects of climate change (e.g. Soberón & Peterson 2004, and references cited therein). As distributed biogeographical and environmental datasets become more available and better integrated, the need for

*Email: john@mcbi.org

simple but flexible tools to exploit them will grow, and the outputs will be extended to more uses. It is vital to understand the nature of the data and the uses to which tools and their outputs can appropriately be put. In this proof of concept study, we explore some characteristics of mapping tools and their output.

The most fundamental biogeographic data concern organism distribution. One convention for depicting distribution is plotting known occurrences as dots (points) on a map. With rare exceptions, these dots are not intended to represent the entire distribution of the taxon in question. A range map is commonly derived from such a dot map, the outermost bounds of a polygon which represents the taxon's distribution connecting the most peripheral dots of the taxon's known occurrence or the points at which organism density falls below a particular threshold (e.g. MacArthur 1972). Such a polygon commonly overestimates the taxon's range. In the marine realm, the range of a shallow-water species occurring throughout the tropical Pacific would cover the entire tropical Pacific Ocean, including the deep water between islands (as in e.g. Fautin & Allen 1992). A more ecologically realistic approach is to correlate actual occurrences with physical, chemical, or biological data (e.g. MacArthur 1972), so, for example, a shallow marine species would be depicted as occurring only around land masses or on banks and shoals.

Thus, more than a collection of geographic coordinates, a range is a manifestation of characteristics of the habitat (biotic and abiotic) that limit or support the organism of interest. A range is inherently a large-scale concept based on observed occurrences; however, range analysis does not necessarily predict organism presence at any specific point. We illustrate some alternative approaches to modeling and understanding habitat distributions for marine organisms by analyzing data from 3 databases with the KGSMapper (Kansas Geological Survey Mapper), an application for interactive analysis of georeferenced occurrence records of marine organisms with gridded environmental data. It is one of a class of electronic tools that, by making it progressively easier to develop correlative analyses from occurrence and environmental data, are rapidly supplanting traditional approaches to interpretive mapping, which tend to be tedious and difficult to replicate. We discuss some issues in evaluating these sorts of analyses. Computer tools and databases cannot substitute completely for knowledge and judgment, however, and the tool we discuss provides ways in which the investigator can interact with and modify the datasets used in order to explore or test hypotheses and tune the nature of the output to the question of interest, rather than simply generating a 'hard-wired' occurrence prediction.

Applications such as WhyWhere (http://biodi.sdsc.edu/ww_home.html) and GARP (www.lifemapper.org/desktopgarp/, http://biodi.sdsc.edu/Doc/GARP/Manual/manual.html), which offer computationally sophisticated approaches to associating environmental and occurrence data (e.g. genetic algorithms), provide the user limited control over datasets and particularly data processing. Tools such as BIOCLIM (http://cres.anu.edu.au/outputs/anuclim/doc/bioclim.html) are confined to or work best in terrestrial habitats. No single approach will be optimal for all questions, or for the needs of all potential users (Fielding & Bell 1997; compare assessments of GARP by Beauvais et al. 2004, Drake & Bossenbroek 2004); in making a choice, consideration must be given to types, scale, quality, and quantity of data available, questions to be addressed, and verifiability of the product (e.g. Fielding & Bell 1997, Manel et al. 2001, Beauvais et al. 2004, Drake & Bossenbroek 2004).

We investigated the issues listed below by generating probabilistic maps of potential habitat occurrence, depicting large-scale areas suitable for survival of these organisms, not organism presence–absence inferences. We used the KGSMapper to analyze the occurrence of habitat suitable for anemonefishes (which may be referred to as clownfishes) and their host sea anemones. The fact that the fishes are obligate symbionts of the anemones (although individual anemones may be found without anemonefish) make this an ideal test case for validating model output: we did not have to go to the field to determine if the organism occurs where we inferred it would, but could compare the inferred occurrence of suitable habitat for the organism to that of the actual occurrence of its symbiont. It is also ideal as a test case in being typical of datasets available for non-fisheries marine species. We discuss model outputs, often termed range, habitat, and niche predictions. Such outputs are commonly used within the natural range of a taxon to guide field work, conservation efforts, etc., and outside the natural range to identify areas vulnerable to biological invasion.

1. Sampling issues. Datasets for a diversity of environmental parameters may be available. The outcome of occurrence predictions or range inferences will be affected by which variables are selected, and how. True niche models (e.g. Peterson 2001, Raxworthy et al. 2003, Soberón & Peterson 2004) must include parameters of the biotic environment beyond strictly habitat characteristics.

2. Data quality. Models must be robust in the presence of questionable or erroneous data points. Particularly for meta-analyses, which use datasets from a variety of sources, the data are likely to vary in accuracy, precision, and resolution, making it unlikely that data quality will uniformly meet the desired standards of

any individual user or application. Therefore, tools are needed for evaluating and/or cleaning datasets when there is a basis for doing so, and the criteria for these actions must be clear.

3. Data quantity. The effect of the number of data on inferences is vital to recognize (e.g. Stockwell & Peterson 2002). A common use of modeling is to infer the biogeographic range of a taxon for which the documented occurrence records almost certainly fall far short of encompassing the actual range. This situation is extremely common for marine invertebrates, particularly for analyses at the species level, but is by no means restricted to them (e.g. Beauvais et al. 2004). Models can provide insight into the areas in which data will be most economical or efficient to sample in order to verify the true extent of the range.

4. Validating or testing results. Assessing predictions or inferences is a desideratum (Fielding & Bell 1997) in this, as in any hypothesis-testing. The end-members on the predictive continuum are a broad-brush approach that minimizes errors of omission and a focused approach that minimizes errors of commission (Fielding & Bell 1997, Anderson et al. 2003). In dealing with the continuum of quality and/or extent inherent in habitat assessment at large spatial scales, omission and commission are not binary no–yes choices, as is typically the case in dealing with presence–absence of organisms; different tests and criteria are called for.

5. Identifying controlling factors. Drake & Bossenbroek (2004, p. 939–940) appealed to scientists to 'develop methods to identify the factors that causally determine species range, and not simply make predictions based on correlations.' Characteristics of a taxon's range or physiology may suggest that particular environmental parameters control its occurrence.

DATA AND METHODS

Data sources and organisms. The organism distribution data for both taxa are georeferenced point occurrences; the third dataset includes gridded coverages of environmental parameters. Having come from 3 proximate providers, all of which compiled data from multiple ultimate sources, our data are unlikely to be homogeneous in quality and scale.

Anemonefishes, which are widespread in the tropical and subtropical Indo–West Pacific but are absent from the eastern Pacific and the Atlantic, occur in nature only with sea anemones of 10 species belonging to 5 genera in 3 families; the fish population is limited by the number of suitable hosts (Fautin & Allen 1992). Anemonefishes belong to 2 genera (*Amphiprion*, with 25 species, and *Premnas*, with 1) in a single subfamily, and vary in host specificity, some associating with only 1 species of host, but most occurring with multiple hosts (Fautin & Allen 1992). Because all host anemones possess photosymbionts, they and their fish symbionts occur only in shallow water (Dunn 1981), typically in waters less than 100 m deep. The distribution of these animals, therefore, is constrained both environmentally and biologically.

In a first approximation, the 10 species of anemones are ecologically similar, and the 26 species of fishes are likewise similar; this allows us to use as our units of taxonomic analysis all host anemones and all anemonefishes. We extracted occurrence data for the anemones from the online resource 'Biogeoinformatics of Hexacorals' (www.kgs.ku.edu/Hexacoral; hereafter referred to as 'Hexacoral'). In the biological database of Hexacoral, which was assembled from the published literature, all names used to refer to a single species are linked, and names that have been applied to more than 1 species are distinguished. Anemonefish occurrence data were downloaded from FishBase (www.fishbase.org), which has been assembled from published records, museum catalogs, and other sources.

The environmental data, also served from Hexacoral, were assembled from public-domain datasets (sources identified in the metadata associated with each dataset) that are global in coverage. Data were gridded in a register at 0.5° resolution (~55 km per side at the equator), which is a typical resolution for global environmental datasets. Datasets with native resolutions other than 0.5° were sampled or aggregated to conform to the grid; for a variable with a native resolution finer than 0.5° (such as the 2′ ETOPO2 bathymetry), within-cell variability and extremes were calculated. Most values are annual or monthly averages. Of the >200 datasets in Hexacoral, 13 especially relevant to anemonefishes and their hosts are currently available for use with KGSMapper; future versions will make the other datasets accessible. In addition to limitations imposed by the size of grid cells, a significant caveat is that the marine datasets used to generate many of the variables typically fail to represent much of the temporal and spatial variability in nearshore environments.

Tools and analytical procedures. KGSMapper is an interactive web-based mapping tool that permits a user to create maps of inferred distribution in a straightforward manner. The basic calculations can be done in a spreadsheet, although much of the power of KGSMapper derives from its ability to display and manipulate the data in a Geographical Information System (GIS) environment. Its flexibility allows a user to select approaches relevant to the goals of the study and to apply expert judgment in editing datasets. It currently uses a tightly integrated environmental database and front end (Oracle 9i RDBMS with Cold Fusion) with, on the server side, ArcIMS web-mapping software (www.esri.com/software/arcgis/arcims/index.html). Occurrence records are plotted in real time on a map through an XML-coded data structure based on the Ocean Biogeographic Information System (OBIS) schema, an extension of Darwin Core 2 (http://iobis.org/obis/obis.xsd). KGSMapper and its associated environmental data are freely available; it is operational through the Hexacoral website (above), the OBIS website (www.iobis.org), and those of some OBIS partners (e.g. CephBase: www.cephbase.org; FishBase: above).

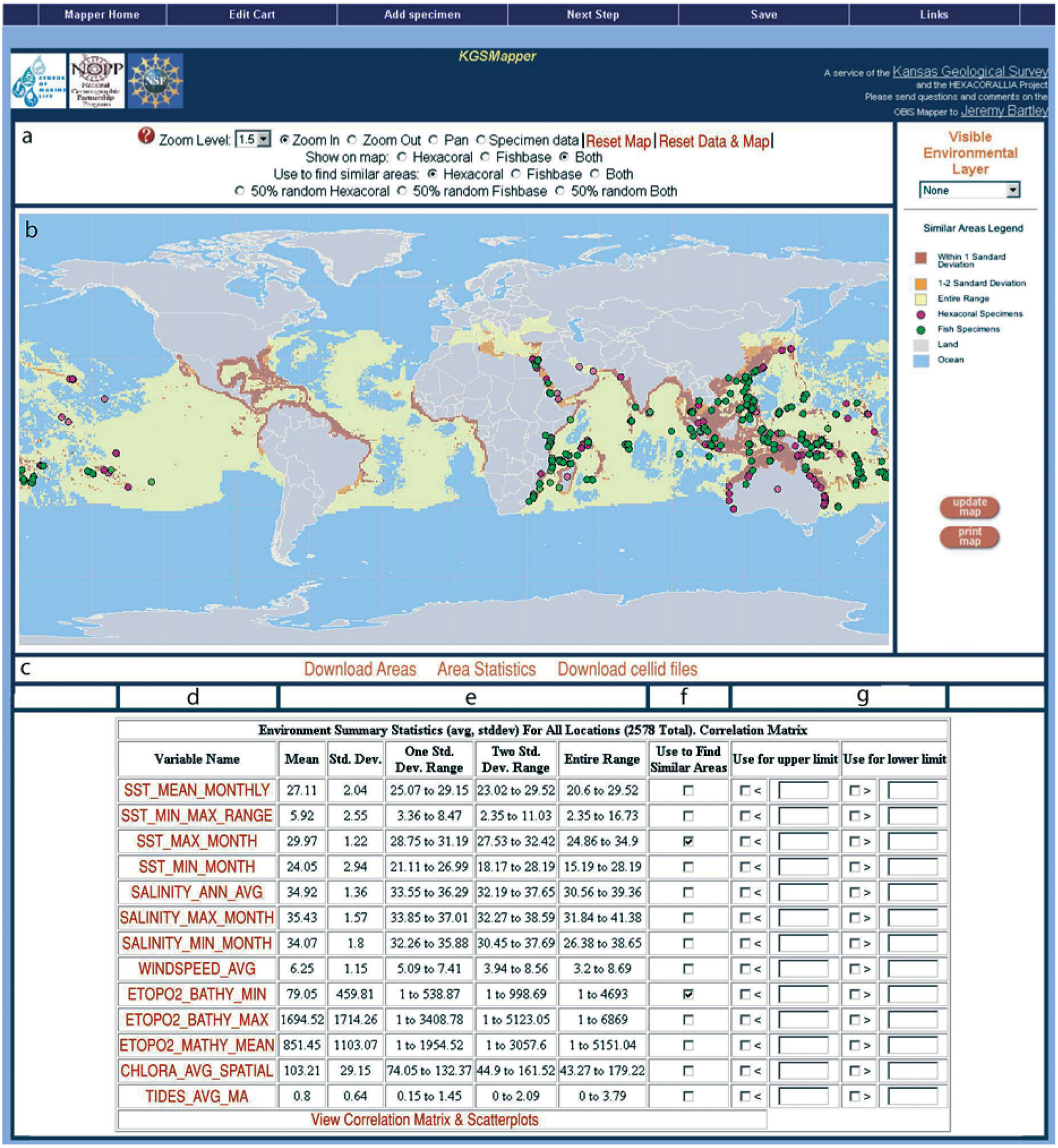
In our analyses, locality records are the 0.5° grid cells containing organism occurrences. Thus, the number of occurrences may not equal the number of locality records in the dataset. Cells with 1 or multiple occurrences are indistinguishable in our analyses — a single occurrence serves to qualify a cell and its environmental variable values as habitat. Conversely, for an occurrence falling on a cell boundary between 2 or among 4 cells, all cells are included in the analysis.

The version of KGSMapper used for this study (http://hercules.kgs.ku.edu/website/specimen_mapper) currently interacts only with the data discussed here. Table 1 summarizes the features of the KGSMapper. Fig. 1 shows the KGSMapper web page; its functions and features are described below, and in the figure caption. KGSMapper plots organism occurrences and provides summary values of 52 environmental variables for all cells in which there is at least 1 occurrence record. Our tests were constrained by the variables available from the main database, and by inherent resolution limitations of working at global scales with primarily marine parameters. These are practical matters — neither is constrained in theory.

Inferences of where suitable habitat occurs for members of the taxon are based on the environment of places where they are known to occur. The user selects the variables by checking the relevant boxes under 'Use to Find Similar Areas' (Fig. 1, Panel f). When the user selects 'Update Map,' KGSMapper builds and executes a query to find the 0.5° cells having all values within 1 standard deviation (SD) of the means of the environmental variables at the occurrence locations, those within 2 SD, and those within the total value range for all selected variables. The results, displayed as an interactive map (Fig. 1), are also available as tabulated statistics (by clicking a link in Fig. 1, Panel c). For 0.5° cells to be classed as within 1 SD, depicted as dull red on the map, all the selected variables must be within 1 SD of the mean of the values of the same variable in cells containing occurrence records. Orange signifies cells in which the value for all selected variables falls within 2 SD of the mean of the values for the selected variable(s), but at least 1 falls beyond 1 SD, and yellow signifies cells beyond 2 SD to the full range of the values known ('outliers'). This probabilistic approach is appropriate in dealing with habitat, which is a continuum from favorable to marginal. It also allows a user to focus attention where habitat or data are optimal, by recalculating a map that eliminates those original cells that have values in the outlier region or beyond 1 SD. This is done by selecting, respectively, the 'Remove All Cells Outside 2 Std. Deviation ranges from cart' or 'Remove All Cells Outside 1 Std. Deviation ranges from cart' options that appear at the bottom of the statistics pop-up page.

Table 1. Features of the KGSMapper tool used for analyses reported here. Last 3 points refer to features still under development

Features
1. Dynamic mapping of selected occurrences
2. Selectable map background
3. Data point identification with link to source database
4. Short list of selectable environmental variables
5. Viewable environmental variable metadata
6. Viewable distribution histogram of individual environmental variables for selected occurrences
7. Correlation matrix of environmental variables for selected occurrences
8. Pairwise scatterplots of environmental variables for selected occurrences
9. Map zoom controls region of analysis, occurrences selected
10. Environmental data table reflects selected locations
11. Range localities classified within 1 SD, 2 SD, and total range of selected environmental variables
12. Downloadable file of occurrences
13. Downloadable shape file of inferred areas
14. Downloadable table of relationships among occurrences, sample cells, and inferred areas
15. Downloadable table of cell IDs for all areas in analysis
16. Eliminating individual records from working dataset
17. Limiting maximum and/or minimum values for environmental variables
18. Comparing or combining 2 datasets
19. Ability to use a random 50% of locations, tested with others
20. User can save and return to a modified dataset
21. User can upload an independent dataset for analysis


Mapper Home
Edit Cart
Add specimen
Next Step
Save
Links
KGSMapper
A service of the Kansas Geological Survey
and the HEXACORALLIA Project
Please send questions and comments on the
OBIS Mapper to Jeremy Bartley
a
b
c
d
e
f
g
Reset Map
Reset Data & Map
Visible Environmental Layer
Similar Areas Legend
update map
print map
Download Areas
Area Statistics
Download cellid files
View Correlation Matrix & Scatterplots

Zoom Level: 1.5 ● Zoom In ○ Zoom Out ○ Pan ○ Specimen data

Show on map: ○ Hexacoral ○ Fishbase ● Both

Use to find similar areas: ● Hexacoral ○ Fishbase ○ Both

○ 50% random Hexacoral ○ 50% random Fishbase ○ 50% random Both

Visible Environmental Layer: None

Similar Areas Legend

- 🟥 Within 1 Sandard Deviation
- 🟧 1-2 Sandard Deviation
- 🟨 Entire Range
- ● Hexacoral Specimens
- ● Fish Specimens
- ⬜ Land
- 🟦 Ocean

Environment Summary Statistics (avg, stddev) For All Locations (2578 Total). Correlation Matrix

Variable Name	Mean	Std. Dev.	One Std. Dev. Range	Two Std. Dev. Range	Entire Range	Use to Find Similar Areas	Use for upper limit		Use for lower limit	
SST_MEAN_MONTHLY	27.11	2.04	25.07 to 29.15	23.02 to 29.52	20.6 to 29.52	□	□ <		□ >	
SST_MIN_MAX_RANGE	5.92	2.55	3.36 to 8.47	2.35 to 11.03	2.35 to 16.73	□	□ <		□ >	
SST_MAX_MONTH	29.97	1.22	28.75 to 31.19	27.53 to 32.42	24.86 to 34.9	☑	□ <		□ >	
SST_MIN_MONTH	24.05	2.94	21.11 to 26.99	18.17 to 28.19	15.19 to 28.19	□	□ <		□ >	
SALINITY_ANN_AVG	34.92	1.36	33.55 to 36.29	32.19 to 37.65	30.56 to 39.36	□	□ <		□ >	
SALINITY_MAX_MONTH	35.43	1.57	33.85 to 37.01	32.27 to 38.59	31.84 to 41.38	□	□ <		□ >	
SALINITY_MIN_MONTH	34.07	1.8	32.26 to 35.88	30.45 to 37.69	26.38 to 38.65	□	□ <		□ >	
WINDSPEED_AVG	6.25	1.15	5.09 to 7.41	3.94 to 8.56	3.2 to 8.69	□	□ <		□ >	
ETOPO2_BATHY_MIN	79.05	459.81	1 to 538.87	1 to 998.69	1 to 4693	☑	□ <		□ >	
ETOPO2_BATHY_MAX	1694.52	1714.26	1 to 3408.78	1 to 5123.05	1 to 6869	□	□ <		□ >	
ETOPO2_MATHY_MEAN	851.45	1103.07	1 to 1954.52	1 to 3057.6	1 to 5151.04	□	□ <		□ >	
CHLORA_AVG_SPATIAL	103.21	29.15	74.05 to 132.37	44.9 to 161.52	43.27 to 179.22	□	□ <		□ >	
TIDES_AVG_MA	0.8	0.64	0.15 to 1.45	0 to 2.09	0 to 3.79	□	□ <		□ >	

Fig. 1. KGSMapper page: the inferred range displayed is based on anemone distributions, maximum monthly sea-surface temperature (SST), and minimum depth value for the grid cells containing anemone occurrences (checked in boxes below map). (a) Zoom and pan controls on top line select region and scale. Clicking a point with ‘Specimen data’ activated produces a pop-up window containing species name(s) and coordinates, values for environmental variables in each cell in the selected area, summary of environmental statistics for all cells containing an occurrence record, and the option of removing the point from the analysis. Second line selects sample points displayed. Third line selects sample set of cells used. Fourth line randomly selects ~50 % of one or both datasets to make a range inference to be tested with the remaining cells. (b) Map shows both datasets with localities distinguished by color of points (purple: sea anemones; green: anemonefishes) and inferred distribution of suitable habitat based on the selected environmental variables (below). Cells in areas colored dull red have values for all variables used for the inference within 1 SD of their means in the record-containing cells, orange is for cells between 1 and 2 SD, and yellow is for the rest of the total range. (c) Links below map provide a download of shapefiles for the areas, a table of statistics of occurrences in both datasets relative to the cells in each range class, or a set of tables of the grid cell identifiers for the cells in each SD category by record contents. (d) Link from the variable name brings up a histogram showing distribution of values and statistics for variable values from the selected locations. Environmental parameters are SST (monthly mean, maximum, minimum, and range), salinity (annual averages, and monthly minimum and maximum), average windspeed, depth (based on ETOPO2 bathymetry: minimum, maximum, and mean), average chlorophyll *a* concentration, and average tidal amplitude. (e) Statistics for each variable reflect the dataset selected by the map display (Panel b). (f) Check boxes for selecting variables with which to ‘update map’ and infer ranges. (g) Check boxes for entering minimum and/or maximum values to restrict the locations selected. Bottom line: link displays a correlation matrix (Table 2) showing linear regression coefficients for each pair of environmental variables, based on values selected in the map display. Values for the correlation coefficients in the matrix cells are linked dynamically to scatterplots of the selected values of each pair of variables. Quality of Figs. 1 to 4 corresponds to that of the images on the computer screen

The menu bar at the top of the page (Fig. 1) provides links to other parts of the site and 2 editing functions. 'Add Specimen' permits a user to augment the occurrence dataset; the 'Edit Cart' link allows a user to eliminate entries from the list of occurrences. The user can also review and edit individual location records with 'Specimen Data' (Fig. 1, Area a). The link 'Next Step' takes the user to the menu of all 200+ environmental datasets in Hexacoral — these currently do not otherwise interact with KGSMapper, but a later version will allow a user to select from all datasets. The 2 right-hand columns (Fig. 1, Area g) allow the user to select upper and lower limits for environmental data, eliminating cells with values outside a specified interval from the analysis. Statistical analyses of both the variables and the inferred ranges can be viewed and downloaded, as can lists of cell IDs and ESRI shapefiles (Fig. 1, Area c). The KGSMapper, which can show 2 groups of taxa concurrently, provides the option to choose which taxa will be displayed (fish, anemones, or both) and/or used as the basis for the range inference (Fig. 1, Area a). In addition, the user can withhold a random selection of ~50% of the records for either dataset or for both datasets, infer a range with the remaining half, and test the product using the withheld records (Fig. 1, Area a).

Because organism occurrences are points (which define the 0.5° cells of analysis), not coverages, inferring the distribution of the habitat suitable for 1 taxon based on distribution records for another differs from inferences using environmental data. Only qualitative matches are possible using maps. A quantitative assessment can be made by determining the number of cells inferred to contain suitable habitat for 1 taxon, based on occurrence records for that taxon, then determining the proportion of known occurrences for the other taxon falling within those cells.

Analyses. We considered the effects of various aspects of the data on model outcome, addressing the issues we raised in the 'Introduction'.

Selection and effects of environmental variables (Issues 1 and 5, see 'Introduction'). We investigated which variables can explain occurrence of the subjects and, if a selection is to be made among them, the basis for choice. We tested 5 variables individually and combined into 4 groups (below). Some of these are known to affect occurrence of anemonefishes and their sea anemone hosts (sea-surface temperature [SST], depth, salinity); others (tidal amplitude and productivity, for which chlorophyll *a* concentration is a proxy) were tested to determine if they might have an effect. We also examined alternative parameters (maximum, minimum, and mean) of some variables (results not reported); minimum SST was chosen because the restriction of the animals to the tropics makes it likely that minimum temperature limits their distribution more than mean or maximum. The correlation matrix (Table 2) assesses the degree to which environmental variables covary within the region selected; this tool permits the investigator to explore the effects of spatial auto-correlation and covariance between variables, in order to help guide variable selection for the question being addressed. The strongest correlations among variables used in this study are within the variants of SST, salinity, and depth; only 1 from each category was used. For example, as might be expected, maximum and mean SST are highly correlated (but minimum SST is less so).

The following groups were selected to determine the effect on output of a number of variables: (1) minimum SST and minimum depth, (2) as Group 1 plus minimum salinity, (3) as Group 2 plus average chlorophyll *a* concentration, and (4) as Group 3 plus tidal amplitude.

Table 2. Correlation matrix of variables for the datapoints associated with fishes and anemones as it appears on screen. 1A: SST_mean_monthly; 1B: SST_min_max_range; 1C: SST_max_month; 1D: SST_min_month; 2A: Salinity_ann_avg; 2B: Salinity_max_month; 2C: Salinity_min_month; 3: Windspeed_avg; 4A: ETOPO2_bathy_min; 4B: ETOPO2_bathy_max; 4C: ETOPO2_bathy_mean; 5: CHLORA_avg_spatial (CHLORA: chlorophyll *a* concentration); 6: Tides_AVG.MA (Tides, Average Maximum Amplitude)

	1A	1B	1C	1D	2A	2B	2C	3	4A	4B	4C	5	6
1A	1	−0.7148	0.791	0.9466	−0.448	−0.3571	−0.2782	−0.703	−0.1105	0.0228	−0.0329	0.0153	−0.0916
1B	−0.7148	1	−0.1629	−0.8921	0.4939	0.486	0.1148	0.56	−0.0222	−0.2255	−0.1698	0.2258	0.1446
1C	0.791	−0.1629	1	0.5912	−0.2069	−0.0804	−0.3012	−0.4896	−0.177	−0.1759	−0.1977	0.2008	−0.0183
1D	0.9466	−0.8921	0.5912	1	−0.4985	−0.4341	−0.2317	−0.6907	−0.0629	0.1038	0.0482	−0.0905	−0.1266
2A	−0.448	0.4939	−0.2069	−0.4985	1	0.9494	0.4293	0.3026	0.0345	0.0938	0.0872	−0.222	−0.1283
2B	−0.3571	0.486	−0.0804	−0.4341	0.9494	1	0.1772	0.1744	−0.0178	−0.0235	−0.0157	−0.0775	−0.1416
2C	−0.2782	0.1148	−0.3012	−0.2317	0.4293	0.1772	1	0.3792	0.1102	0.2677	0.2382	−0.346	0.0042
3	−0.703	0.56	−0.4896	−0.6907	0.3026	0.1744	0.3792	1	0.1233	0.1369	0.1718	−0.2946	−0.0884
4A	−0.1105	−0.0222	−0.177	−0.0629	0.0345	−0.0178	0.1102	0.1233	1	0.5028	0.7361	−0.2887	−0.1417
4B	0.0228	−0.2255	−0.1759	0.1038	0.0938	−0.0235	0.2677	0.1369	0.5028	1	0.9033	−0.6949	−0.2789
4C	−0.0329	−0.1698	−0.1977	0.0482	0.0872	−0.0157	0.2382	0.1718	0.7361	0.9033	1	−0.6195	−0.2404
5	0.0153	0.2258	0.2008	−0.0905	−0.222	−0.0775	−0.346	−0.2946	−0.2887	−0.6949	−0.6195	1	0.3395
6	−0.0916	0.1446	−0.0183	−0.1266	−0.1283	−0.1416	0.0042	−0.0884	−0.1417	−0.2789	−0.2404	0.3395	1

Uncertain data quality (Issue 2). Both organism datasets contain locations known to be inaccurate (of course, we cannot know if there are additional inaccurate locations). Inaccurate locations can sometimes be identified by their associated depth; because the anemonefishes are constrained to live within the photic zone (operationally to ~100 m) by the photosymbionts of the host anemones, depths greater than 100 m strongly suggest an erroneous location. We inferred potential habitats of both fish and anemones, with and without eliminating cells, at minimum depths of >100 m.

Validating or testing range inferences (Issue 4), including making inferences about the effects of data quantity (Issue 3). We compared the outcomes of inferring habitat of each taxon based on records of another, and inferring the habitat of each taxon based on ~50 % of the records for a taxon selected randomly by the KGSMapper tool. We also demonstrated the effects of eliminating from the initial dataset points in cells with values for environmental parameters >1 SD and >2 SD.

In this case study dealing with the continuum of quality and/or extent inherent in habitat assessment, KGSMapper output ranks probability of matching habitat characteristics rather than a dichotomous occurrence or not of organisms; for this reason and because assessment of known absences at the scales used (global extents and ~2500 km^2 grid cells) are impractical, output cannot be evaluated by confusion matrix measures (Fielding & Bell 1997, Manel et al. 2001). We evaluated output by what we term 'effectiveness' and 'efficiency,' assessing the distribution of cells among the intervals 0 to 1 SD, >1 to 2 SD, and >2 SD. The assumption, as in most habitat models, is that the distribution of cells inferred to contain suitable habitat will reflect that of occurrence-containing cells. For each interval i, the number of cells containing an occurrence is a_i, and the number of cells within the range is n_i. a_T is the total number of cells containing an occurrence record over n_T (the total of cells over all n_i). 'Effectiveness' is the ratio a_i/a_T—for each interval, the fraction of occurrences contained within the cells of that interval; a high value indicates inclusiveness or relative lack of false negatives. 'Efficiency' is the fraction of total occurrences per area (number of cells) inferred; we use the ratio $(a_i/a_T)/n_i$. This represents the density of positive occurrences; increasing values indicate a decrease in false positives. Effectiveness and efficiency, which are related but not identical to the confusion matrix measures of predictive power, sensitivity and prevalence, function within a run of the model; effectiveness minimizes errors of omission, and efficiency minimizes errors of commission. The data selection and editing tools permit the ratio of efficiency to effectiveness to be adjusted according to the questions and data of interest; like the output itself, evaluation of the results will necessarily be application specific.

RESULTS

Environmental variables

For each set of environmental variables, we did 3 analyses, 1 for each group of organisms individually and 1 for the 2 together. We illustrate examples of inferring the distribution of suitable habitat for each combination. Of datasets in the KGSMapper, the parameters of chlorophyll *a* concentration (Fig. 2a), minimum salinity (Fig. 2b), and tidal amplitude and wind speed (not shown) did not discriminate suitable habitat at the geographic scale of this analysis. Combinations of 2 or more of these variables provided no more resolution than any single variable analyzed individually. SST discriminated best for the habitat of these organisms latitudinally, with results differing somewhat depending on the parameter used (compare Fig. 2c,d for maximum and minimum monthly SST, respectively). Two approaches were tried to consider depth, which also restricts distribution of these animals: Fig. 2e resulted from using occurrence data alone, whereas Fig. 2f excluded the cells with minimum depths of >100 m. The number of cells inferred to contain suitable habitat (total range) was reduced by >85 % as a result of editing for depth (Table 3, Fig. 2e,f). The outlier

Table 3. Inferences of suitable habitat using minimum SST and minimum depth as environmental variables, and occurrence data. Edited inferences (right-hand column for each taxon) used only records in cells in which minimum depth was <100 m. The line '0–2 SD' is the total of the preceding 2 lines. Ctot: total number of cells inferred to contain suitable habitat; Crec: number of record-containing cells; Rec: number of occurrence records; n = 641 for anemones; n = 1937 for fish

	Anemones						Fish					
	Unedited			Minimum depth <100 m			Unedited			Minimum depth <100 m		
	Ctot	Crec	Rec	Ctot	Crec	Rec	Ctot	Crec	Rec	Ctot	Crec	Rec
0–1 SD	6187	261	385	5331	244	385	7661	250	1281	4791	221	1211
1–2 SD	3450	103	207	1853	90	188	9150	119	538	1719	88	492
0–2 SD	9637	364	592	7184	334	573	16811	369	1819	6510	309	1703
>2 SD	49142	63	35	915	6	30	39379	58	107	731	27	20
Total range	58779	427	627	8099	340	603	56190	427	1926	7241	336	1723

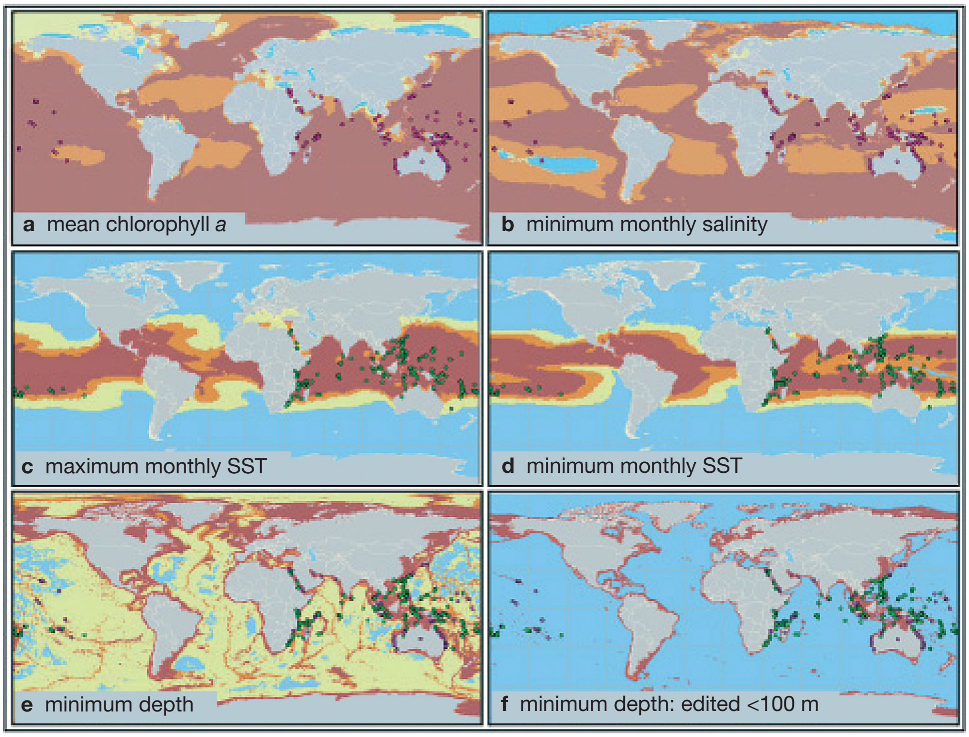


Fig. 2. Suitable habitat inferred on the basis of single variables and organism distributions. Habitat suitable for anemones inferred from anemone occurrences is based on values from the cells containing occurrence records: (a) chlorophyll *a* concentration and (b) minimum monthly salinity. Habitat suitable for fish inferred from fish occurrences: (c) maximum monthly SST and (d) minimum monthly SST. Combined fish and anemone habitat inferences: (e) minimum depth and (f) minimum depth excluding cells with values >100 m. Chlorophyll and salinity do not account for habitat, individually or in combination; similar results were obtained with tidal amplitude and wind speed (not shown). The depth constraints and the latitudinal controls imposed by SST provide a powerful combination (see Figs. 3 & 4)

category (>2 SD) was most heavily affected for anemones; editing reduced the number of 0 to 1 SD cells by 14% and of 0 to 2 SD cells by 25% in the case of anemones. The figures for fishes were 37 and 61%, respectively.

Fig. 3 illustrates the way datasets can be cleaned or edited based on either specific knowledge or statistical evaluation; to allow details to be seen clearly, they show only the part of the world where most species of these animals occur, but the analyses which led to these results made use of global data. Fig. 3a,b shows inferences of anemone and fish habitat, respectively, based on minimum depth and minimum SST, which individually provided reasonable first approximations to defining appropriate habitat (above). Fig. 3c,d shows the improvements in both inferred ranges generated by eliminating cells with a minimum depths of >100 m. Fig. 3e,f has been remapped after elimination of all cells >2 SD in Fig. 2a,b. Fig. 3g,h shows the effects of removing all cells >1 SD from the datasets used in Fig. 3a,b. This rigorous cleaning shrinks the geographic range noticeably, but the 0 to 1 and 0 to 2 SD intervals remain relatively similar throughout.

Occurrence data quantity and quality

After removal of 2 anemone localities in the Mediterranean Sea that were clearly due to misidentification of specimens, misapplication of a name, or misstatement of provenance, the datasets contained 641 anemone and 1937 fish records. They included some suspect data points and some of low precision; we retained all to provide a realistic test of habitat inference using the sort of data likely to be available for analysis of non-fisheries species.

Four anemone and 9 fish records fell on land outside a coastal cell; because marine variables are not associated with inland cells, these points were ignored in the analyses. Points on land in a coastal cell were analyzed using the marine variables associated with that cell. Some records on land do not reflect errors: the anemone dataset (for which a precision value is assigned to each georeferenced point) contains low-precision records assigned by a convention that plots the locality in the center of a country or region given as the only location information in the original publication. This results in points on land

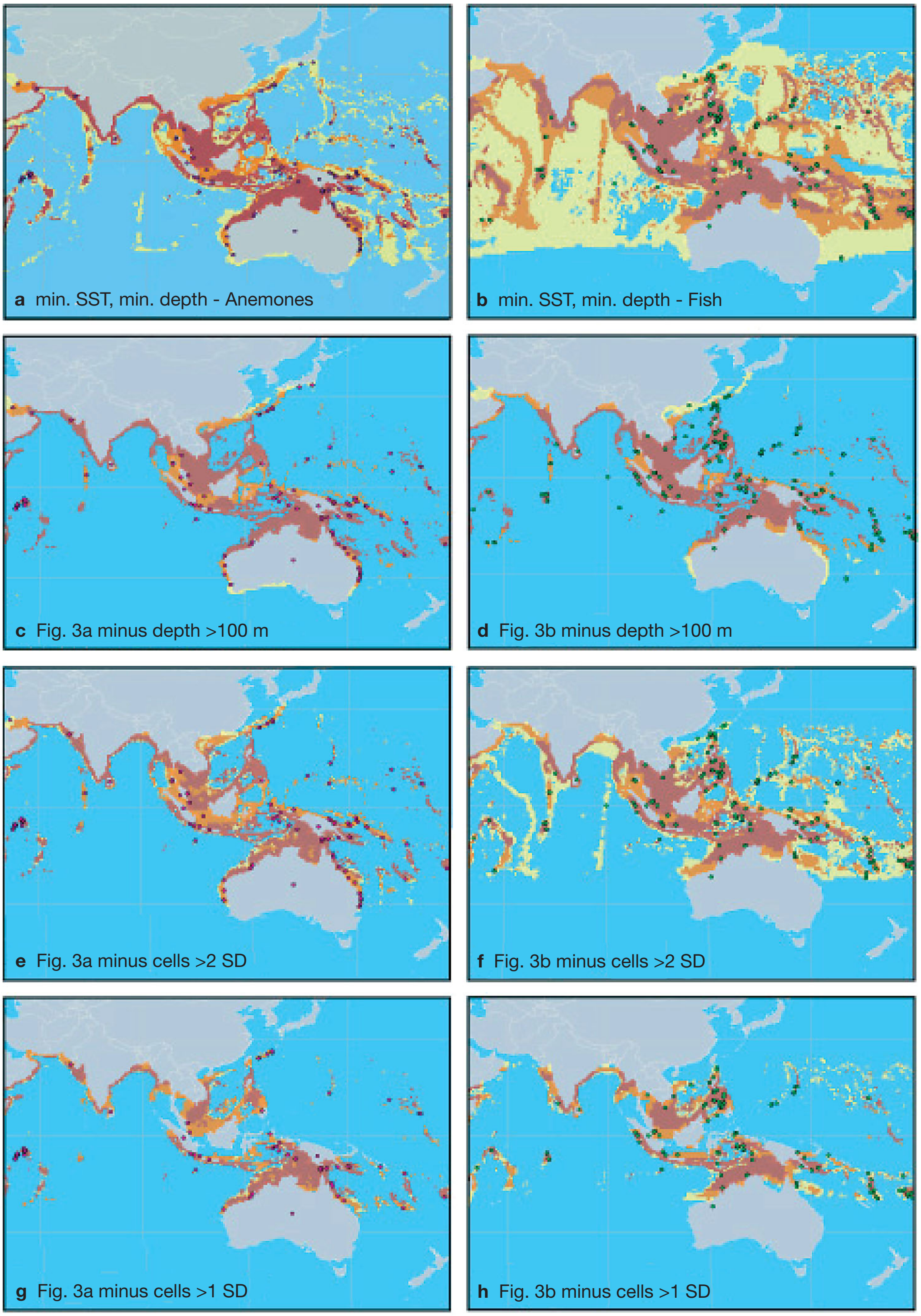

a min. SST, min. depth - Anemones
b min. SST, min. depth - Fish
c Fig. 3a minus depth >100 m
d Fig. 3b minus depth >100 m
e Fig. 3a minus cells >2 SD
f Fig. 3b minus cells >2 SD
g Fig. 3a minus cells >1 SD
h Fig. 3b minus cells >1 SD

Fig. 3. Dataset clean-up and editing features displaying zoomed views of the Australasian region after inferring ranges based on the global dataset. (a) Habitat suitable for host anemones based on anemone occurrences and (b) anemonefish habitat based on fish occurrence records using unedited datasets, with minimum monthly SST and minimum depth. (c), (d) as (a) and (b), respectively, but with datasets edited to eliminate cells having minimum depths of >100 m (see Fig. 1g). (e), (f) as (a) and (b), respectively, but recalculated after eliminating cells in the >2 SD category in the initial analysis. (g), (h) as (a) and (b), respectively, but recalculated after eliminating cells >1 SD in the original analysis. The datasets can also be edited point by point, if desired (Fig. 1, Area a)

Table 4. Data using 50% of anemone-containing cells and minimum SST and minimum depth to infer habitat suitable for the remaining anemones. After editing to exclude cells with depths >100 m, 136 cells were used in this analysis

Trial	No./% cells used for inference	No./% remaining records inferred
1	57/41.91	76/96.20
2	63/46.32	73/100
3	74/54.41	61/98.39
4	64/47.06	69/95.83
5	65/47.79	71/100
6	69/50.74	67/100
7	63/46.32	72/98.63
8	67/49.26	68/98.55
9	62/45.59	71/95.95
10	65/47.79	71/100
11	57/41.91	74/93.67
12	66/48.53	70/100
13	65/47.79	71/100
14	68/50.00	68/100
15	73/53.68	63/100
16	71/52.21	65/100
17	75/55.15	61/100
18	62/45.59	74/100
19	66/48.53	69/98.57
20	71/52.21	62/95.38
Average	66.15/48.64	68.80/98.56
SD	5.01/3.68	4.44/2.01

(e.g. the centroid of Australia for localities given as ‘Australia’), and over water far deeper than that in which these anemones live (e.g. the center point of Fiji for localities given as ‘Fiji’).

Editing to eliminate occurrence-containing cells with minimum depths of >100 m reduced anemone records by ~4% (24) and fish records by ~11% (203), but, because one 0.5° cell may contain >1 occurrence record of a fish or anemone, the number of record-containing cells was reduced by ~20 and ~21%, respectively (Table 3).

Cross-comparison and validation

Areas of suitable habitat for anemones, as inferred using 50% of anemone-containing cells, included between 93.7 and 100% of the remaining known occurrences (Table 4)—as well as many places where the anemones are not recorded as living. Clearly, the best test of our model output would be to seek the animals in places where suitable habitat is inferred to exist and the animals are not known to occur. That being impractical, we ran an analysis using KGSMapper, appropriate environmental parameters, and the native distribution (from FishBase) of anemonefishes.

On a map, known fish occurrences fell largely within areas of inferred habitat suitable for anemones and vice versa. In a quantitative assessment, using minimum SST and minimum depth (see Table 5), areas inferred by anemone occurrences included virtually all places fish are known to occur, a result somewhat improved by editing both datasets for depth. Fish occurrences were less effective in identifying areas suitable for anemones, and editing had little effect (Fig. 3). Thus, at the scale of this analysis, suitable habitat is inferred not to occur where it does not occur (at high latitude and at depth).

To explore the effects of number of environmental variables on inferred ranges, we used the 4 groups of environmental variables listed in ‘Data and methods.’ Fig. 4a,b shows the number of cells within each interval (the former for raw data, the latter for data edited to exclude cells with minimum depths of >100 m), Fig. 4c, d shows effectiveness, and Fig. 4e,f shows efficiency. As single variables were added, effectiveness of the output in the 0 to 1 SD interval declined. However, efficiency increased because the inferred number of cells (n_{0-1}) decreased more rapidly than the number of occurrence-containing cells (a_{0-1}). We found the same pattern within groups of related variables—inferences using maximum or minimum SST plus minimum depth and maximum or minimum SST plus the 4 variables used to generate Fig. 4d indicate that the use of maximum SST is more effective than minimum SST, which is somewhat more efficient than maximum SST.

DISCUSSION

Environmental variables

Individually, the variables of minimum salinity, chlorophyll *a* concentration, tidal amplitude, and wind speed do not identify the occurrence of habitat suitable for anemonefishes and sea anemones (Fig. 2a,b): much

Table 5. Using minimum SST and minimum depth plus occurrence of a symbiotic partner to infer occurrence of habitat suitable for another symbiotic partner, as evaluated by the percentage of target organism occurrences in the various categories of inferred habitat cells (anemones were used to infer fish habitat and vice versa). Unedited inferences used all data; edited inferences eliminated records in cells in which minimum depth was >100 m

Category	Fish habitat inferred from anemones (%)		Anemone habitat inferred from fish (%)	
	Unedited	Edited	Unedited	Edited
0–1 SD	69.4	74.0	53.4	52.4
1–2 SD	24.4	25.9	27.6	28.0
>2 SD	5.9	0.1	14.2	14.6
Total range	99.7	100	94.2	95.0

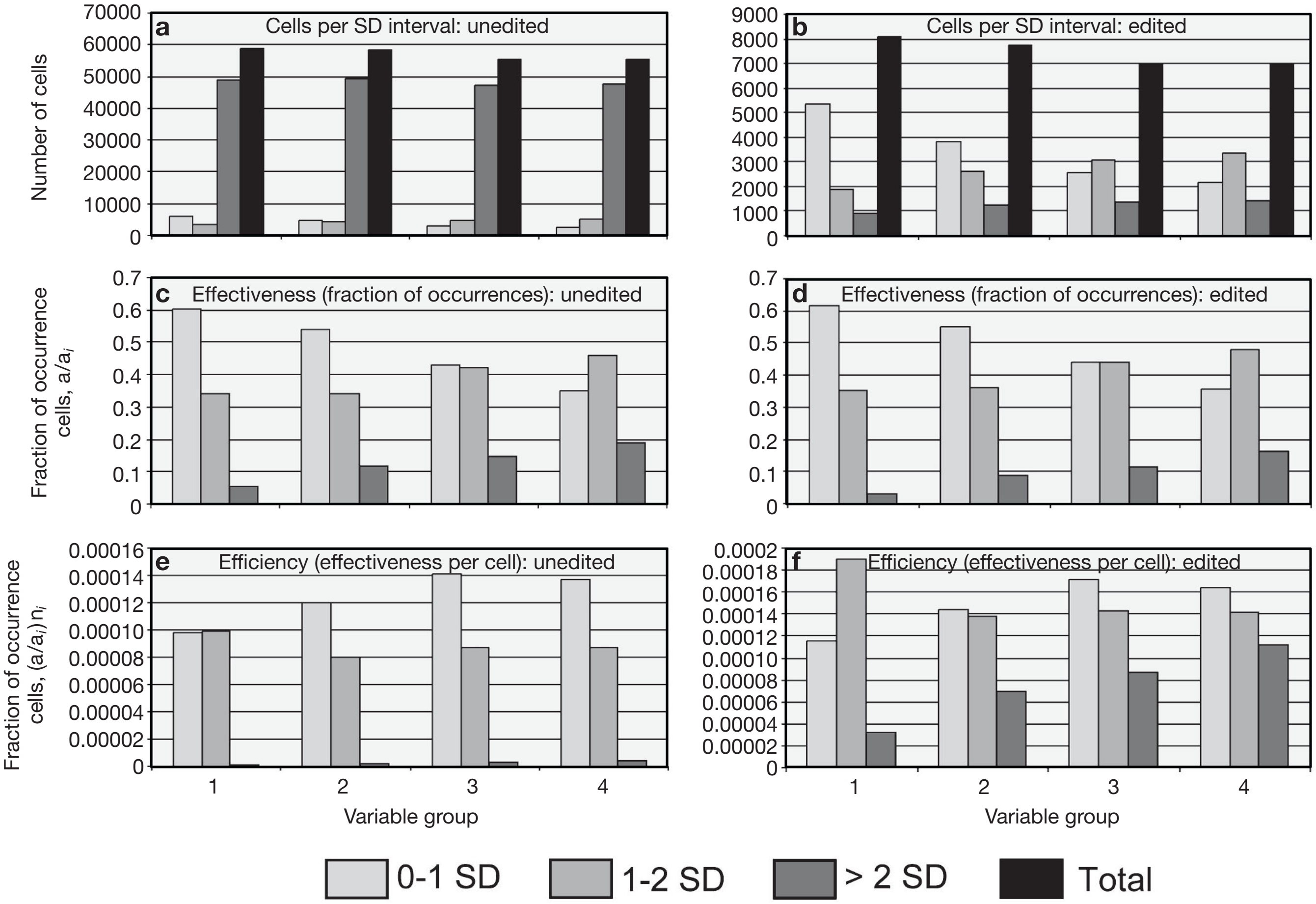


Fig. 4. Distribution of cells, effectiveness, and efficiency of habitat inference for sea anemones as functions of kind and number of variables. (a,c,e) Use values from all data. (b,d,f) Use data edited to exclude cells having minimum depths of >100 m. Numbers on abscissa are variable groups listed in the 'Data and methods' section. (a) and (b) show number of cells; (c) and (d) show effectiveness; (e) and (f) show efficiency

of the ocean has values equal to those of waters in which these animals occur. Although low salinity is negatively associated with anemone occurrence (most sea anemones, including the species that host anemonefishes, are stenohaline; Shick 1991), the resolution of our datasets both temporally (monthly averages) and spatially (0.5° cells based mainly on oceanic measurements) is too coarse to capture its effect. Similar arguments can be made for chlorophyll *a* and for the energy- and exchange-related tide and wind variables. Further, tidal amplitude is unlikely to exert systematic control because it is the relative, rather than absolute, position to low tide that affects anemone survival.

Even highly correlated parameters (Table 2) may not have the same effect. For single variables, maximum and minimum SST (Fig. 2c,d, respectively) infer somewhat different distributions of suitable habitat overall, and in the intervals 0 to 1, 1 to 2, and >2 SD. This is also true in combination with other parameters.

Adding parameters sequentially to minimum temperature and depth (Fig. 4) did not provide increasingly good inferences, from which we conclude that more variables are not necessarily better (cf. Stockwell & Peterson 2002). Quality of the variables, as judged by relevance to occurrence of the taxon in question (Fielding & Bell 1997), seems more important than the number of variables. Quality can be improved by basing the output on values that do not include the outliers (>2 SD) or >1 SD (Fig. 3).

Even with the limited number of environmental variables available in the KGSMapper, choosing variables expected to be relevant to the distribution of any taxon requires some expert judgment, as does determining which relevant variables to use for a given purpose. For example, although maximum SST is quantitatively more effective than minimum SST, it identifies a larger range overall and overextends the northern extent of the fish distribution (Fig. 1, map result). KGSMapper can help to reveal which parameters are most closely correlated with occurrence, and thus may be important in controlling, or describing, distribution.

Occurrence data quality and quantity

The linkage of taxonomic synonyms allows Hexacoral to map occurrences for the species rather than for the name; this also helps to increase the number of records for a species. Thus, rather than synonymous names being viewed as a problem (Soberón & Peterson 2004), if handled appropriately, they can serve to enhance data quantity and taxonomic quality.

The 2 Mediterranean records we removed illustrate the need for expert judgment in selecting both occurrence and environmental data. Machine algorithms that cleanse datasets by purging records from areas well beyond known occurrences risk removing information on range extensions or invasions. An expert may be able to differentiate among potential sources of error by considering date, similar species, taxonomic history of a name, etc., to make suspect records useful, and thereby improve data quality and quantity.

Using only cells with minimum depth values <100 m resulted in a more precisely defined range (Fig. 2f) than merely selecting minimum depth as a variable (Fig. 2e), presumably because some actual occurrences fall in cells with minimum depths of >100 m, due to either error or convention (such as using the center point of the Fiji Islands for all localities given only as 'Fiji' in the anemone dataset).

Such editing for a feature relevant to organism distribution provides a crude assay of data quality. Editing produced a less dramatic change for anemones than for fishes; compare Fig. 3a,c,e,g with Fig. 3b,d,f,h, respectively. This finding is concordant with what is known of the data sources: the anemone records were assembled as a single project (by D. G. Fautin) and have been extensively checked, whereas the fish records are from multiple sources with unknown and diverse authentication procedures. Thus, KGSMapper deals with suspected, inferred, or known erroneous data to provide a justifiable way to limit consideration to reasonable habitat possibilities. By doing so, it inferentially takes absences into account.

It is commonly thought that more environmental variables will improve the sensitivity or precision of a prediction. Fielding & Bell (1997) call this into question in their discussion of the issues of inappropriate variables, the 'costs' of misclassification, and the contexts in which predictive models are evaluated. Fig. 4 illustrates how choice and number of environmental variables affect output in our study system. As we added single variables to the analysis, the number of cells identified as containing suitable habitat declined by ~10%, but it would be a mistake to interpret this as increasing precision; the fraction within 1 SD declined by ~40% in both edited and unedited analyses. KGSMapper statistics are calculated in a univariate manner; as variables are added, the probability declines that any cell will contain values within 1 SD for all of them. Thus, adding a variable that would be expected, based on biology and analysis, to have little control over organism occurrence can eliminate cells that contain suitable habitat—a high price to pay for minimal return in terms of genuinely improved results.

Others have also found that quality of prediction is not necessarily improved by quantity of data. 'Accuracy' of 4 modeling methods, including GARP, used by Stockwell & Peterson (2002) did not increase beyond about 20 data points, 10 producing 59 to 64% 'accuracy' (90% of potential achievement rate using with their methods). Beauvais et al. (2004) achieved 'validation success' rates of 40.0 to 88.2%, the lowest with a dataset of 18 records, another dataset of 20 records had a rate of 80.0%. The effects of geographic scale and habitat heterogeneity on quality of model output have not been addressed formally, but, based on what is now known, this is an issue which should be addressed. The methods of Stockwell & Peterson (2002, p 11) modeled 'widespread species … less accurately'; Raxworthy et al. (2003) achieved a similar result using GARP. Attention must be paid to this subject for marine species, many of which have larger geographic ranges than is typical of terrestrial species for which predictive algorithms were developed (the animals we studied range through about 180° of longitude and 50° of latitude) and occur in 3 dimensions. In one of the few published modeling studies for the distribution of marine species (fish living in the central western Atlantic), Wiley et al. (2003, p 124) also found that, using GARP, results for widespread species were 'weak.'

In addition to large numbers of points, a desideratum for this sort of analysis is independence of data (Fielding & Bell 1997). However, many of the anemone records we used came from a small number of areas and/or investigators; we have found that records for other poorly studied marine organisms may not be truly independent.

Validating or testing results

Use of training data for assessing quality of model output is a common practice (e.g. Anderson et al. 2003). Such data may constitute a portion of known occurrences (e.g. 50% in Peterson et al. 2002, 75% in Beauvais et al. 2004) or areas of occurrence (e.g.

states in Peterson 2001). KGSMapper has a tool that randomly selects ~50% of reported occurrences and uses the associated locality records (grid cells) to infer the remainder of the localities and their associated occurrences (Fig. 1, Area a). If grid cells are the basis for analysis when using a gridded environmental database and a one-to-one relationship between cells and occurrence records does not exist, the use of occurrence records will not be reliable. A random sample of (e.g. 50%) of the locality cells may contain far more than the stated proportion of the sample occurrences (up to 75–80% in 50% of the tests we conducted with anemone data). This greatly increases the apparent quality of the results and is misleading if that level of performance is ascribed to 50% of the occurrences.

A drawback involved in withholding some records as training data is that 'the algorithm cannot take advantage of all known locality records' (Anderson et al. 2003, p 213). The symbiosis allowed us both to use all data and to implement the desideratum of incorporating interspecific information into the model (Fielding & Bell 1997); we used records of 1 organism to infer areas of suitable habitat for another. We ascribe the asymmetry in our results (Table 5) to that in the relationship—although an anemonefish never occurs without an anemone in nature, individual anemones may occur without fish in some areas. Thus, anemone data will somewhat overestimate suitable habitat for fish. This result is consistent with the potential problem in modeling pointed out by Fielding & Bell (1997) of undersaturation of habitat. Accordingly, saturated symbiotic systems such as this case should be particularly favorable as tests of habitat models.

As an indirect assessment of KGSMapper, we used environmental variables from Hexacoral with occurrence data from FishBase for the tropical Indo-Pacific lionfish *Pterois volitans*. The inferred distribution of suitable habitat resembles that of anemonefishes, and includes the coast of the southeastern United States, where it has recently established viable populations (e.g. Semmens et al. 2004).

The addition of environmental variables that do not, and are not expected to, have any real explanatory power has the effect of increasing the apparent efficiency of the range inference. This is an artifact of constraining the basis on which cells are selected, whether or not that constraint has anything to do with organism occurrence. For a group of organisms that has been extensively sampled over most of its range, this will have little effect other than to distort the apparent quality of the range inference. However, for sparsely sampled organisms, such as most marine organisms, inclusion of gratuitous variables could significantly alter the inferred range.

Although we can readily envision application of KGSMapper to dichotomous problems, the analyses presented here cannot be usefully evaluated by confusion matrix methods (Manel et al. 2001), because of the unavailability of useful absence data at the scale of interest. A half-degree grid cell can be as large as 3000 km^2 in area; the organisms of interest range from a few cm^2 to about 1 m^2 in area, and habitat patches may be <100 m^2. The grid cell is best treated as a mosaic of potential habitats, ranging from favorable to stressful to impossible. To provide some assessment of the quality and characteristics of the inferences, we use efficiency and effectiveness, which allow a user to tune the results for a particular purpose based on the relative importance or cost (Fielding & Bell 1997) assigned to errors of commission and omission. For example, a user planning an expedition to sample particular taxa or to devise a scheme for protected areas would probably want to emphasize efficiency (i.e. maximize the probability of finding organisms per unit area covered), while a study concerned with invasion potential, marginal habitats, or range limits would need the most effective (complete) inventory of potential habitat. Moreover, such analysis allows a user to allocate effort where it will most enhance a product of prediction—adding occurrences would improve the product more than adding environmental features. Similarly, Graybeal (1998) found that adding taxa improved resolution of phylogenetic trees more than adding characters.

Maps and model outputs

An occurrence (or dot) map plots localities where members of the taxon have been documented (for example, Fig. 1, Area b, without the inferred areas of occurrence); subdividing occurrences temporally allows comparing distributions through time. An inference about where members of the taxon may occur beyond the known occurrences constitutes a range map. This, too, may be temporally defined, showing, for example, where organisms formerly occurred, but do not occur currently. It may consist of discontinuous patches, as for the anemonefish and their host anemones around land masses. When drawn up the 'old-fashioned way,' a range map is a simple abstraction of occurrences, an inference of where members of the taxon may occur within the same geographical region. A map generated electronically by a tool such as KGSMapper, by correlating environmental parameters with known distributions, is essentially a habitat map, plotting places compatible with the life of the organism of interest.

A habitat map may contain areas of 2 types, and we advocate that these be distinguished from one another. Areas on a habitat map that fall within the broad ambit of the taxon constitute, as defined above, a range map. Such maps are useful for planning, e.g. field research and conservation strategies in that, by depicting realized habitat, they provide reasonable precise inferences about where members of a taxon may actually live. Some habitat maps include areas that fall well outside the known distribution of the taxon, as illustrated in Fig. 1: anemonefishes and their hosts only occur naturally in the Indo–West Pacific, but ostensibly suitable habitat for them occurs in some areas of the Atlantic (especially the Caribbean) and the eastern Pacific. Such a map depicts potential habitat, which is ideal for identifying places vulnerable to invasion. Because the word 'prediction' literally refers to the future, it is appropriately used for areas outside the natural range — that is for areas subject to invasion. Within the general geographical area in which members of a taxon are known to occur, where direct evidence of their occurrence may currently be absent, a model actually infers — rather than predicts — appropriate habitat.

Some model outputs are said to be niche maps; whereas a habitat is defined on the basis of abiotic parameters, a niche also includes biotic parameters (e.g. Peterson 2001, Anderson et al. 2003). Including explicit biotic information in automated tools such as KGSMapper is difficult, because such information is rarely in the form of coverages. The 1 biotic parameter common in oceanographic data is chlorophyll *a* concentration, but this lacks discriminatory value for the occurrence of most organisms such as those we studied (Fig. 2a). We found that, although appropriate habitat for anemonefishes exists outside the Indo–West Pacific, when we included a vital component of the animal's biotic environment, a host anemone, those areas were no longer identified as habitable. We therefore advocate that such relevant biotic factors be explicitly incorporated into models if they are to be considered niche models. In this case we used symbionts, some pairs of which are mutualistic, precisely because this provided a clearly relevant biotic factor with which to test model output. The relevant biotic factors in other analyses may be less obvious.

Thus, anemonefishes are less likely than lionfish to establish viable populations in the coastal southern United States: although abiotic attributes of the habitat, such as temperature and depth, appear suitable for anemonefish existence, anemones that naturally host anemonefishes do not occur there (Fautin & Allen 1992). One way to infer absence is to eliminate deep water cells (cells in which minimum depth is >100 m). A second way to infer absence is to eliminate all fish habitat cells outside the Indo–West Pacific. This is justifiable based on the absence of an obligate symbiont. By contrast, the potential for Hawaii to be invaded by anemonefish is real, because 1 species of host anemone occurs in Hawaiian waters (Fautin & Allen 1992). On the other hand, for species of these anemones that can live in nature without fish symbionts (most of them), we infer that the suitable habitat outside the Indo–West Pacific is vulnerable to invasion. Once individuals of a host anemone are present in a non-native place, they might be follwed by the species of fish able to live with this particular species of host anemone.

Modeling tools

It is difficult and/or impractical to control quality when using merged, distributed datasets. Therefore, analytical and predictive tools must have features that ensure robust output in the presence of questionable data and that offer the user ways to modify the datasets and to assess the results — by improving data quality, by testing hypotheses derived from them, or both. We have shown that the number and distribution of outlier points is an indicator of both the quality of occurrence data and the relevance of the environmental variables selected. Thus, an output that segregates results into categories of diminishing accuracy allows a user to select appropriate subsets of the output. User decisions can be based on the level of data confidence and the purposes for which the output is to be used. With KGSMapper, for example, we found the 1 SD range to be a robust initial estimate of range, even in noisy datasets.

Beyond this passive evaluative approach, KGSMapper has data-editing features that are broadly useful in assessment and research. A user can edit occurrence data: (1) point by point on the map or data list, (2) by geographic area (using the zoom control), (3) by taxon, and/or (4) by editing environmental variables. Future versions of the KGSMapper will have more versatile means of selecting geographic extent and will include explicit absence as well as presence data. In addition to selecting geographic extent, the ability to edit variables provides a means of exercising expert judgment by cleaning the datasets of points that do not conform to relevant environmental controls and of refining the geographic limits of potential ranges; these are ways to incorporate knowledge of absence. An important means of improving the precision of the habitat inferences is provided by allowing a user to remove records that fall beyond a predetermined statistical limit. The user can then recalculate the model with the remaining cells. KGSMapper allows application of expert judgment at both input and output ends of the process; algorithms such as GARP apply it only at the output end (e.g. Anderson et al. 2003, Drake & Bossenbroek 2004).

KGSMapper outputs go beyond simple map visualizations, providing statistical analyses of the individual variables, of the relationships among the variables, and of the occurrence–environment relationships. In addition to allowing analyses in a manipulative GIS environment, KGSMapper has options that permit dynamic data assessment, which enables the user to identify covarying parameters, variables to be edited, and specific ranges of values to be included or excluded.

Models of organism occurrence may contain 2 types of errors: predicting the organism will occur where it does not (false positive, commission, or overprediction) and not predicting the organism to occur where it does (false negative, omission, or underprediction) (e.g. Fielding & Bell 1997, Anderson et al. 2003). Unlike many algorithms, the objective of KGSMapper is to infer the locations of habitat suitable for occurrence of organisms, not organism occurrence itself. Finding the organisms in the habitat clearly demonstrates it is suitable; not finding them, termed by Anderson et al. (2003) 'apparent commission error,' is due to well-known contingencies in occurrence. To regard prediction of habitat in a place that has not been searched as a false positive is to imply perfect knowledge of organism occurrence. Selecting areas for fieldwork is a potential use of the output of such modeling, particularly for poorly sampled taxa; overestimation of habitat occurrence is therefore neither unexpected nor necessarily undesirable. Moreover, a model that identifies all, but only, the places of known occurrence would be tautologous.

Acknowledgements. Financial support was provided by US National Science Foundation Grants OCE 00-03970 (through the National Oceanographic Partnership Program), DEB 99-78106 (in the PEET program: Partnerships for Enhancing Expertise in Taxonomy), and DEB 95-21819 (PEET). Collaborations with LOICZ (Land–Ocean Interactions in the Coastal Zone, an IGBP project) and FishBase (www.fishbase.org), an element of OBIS (the Ocean Biogeographic Information System), were central to this effort. Additional gratitude is due to the US Fish and Wildlife Service; the US National Biological Information Infrastructure (NBII), especially T. D. Beard and M. Fornwall; K. Look and K. Nelson of the Kansas Geological Survey; A. Ardelean, and G. P. Beauvais, University of Wyoming, USA.

LITERATURE CITED

Anderson RP, Lew D, Peterson AT (2003) Evaluating predictive models of species' distributions: criteria for selecting optimal models. Ecol Model 162:211–232

Beauvais GP, Keinath D, Thurston R (2004) Predictive range maps for 5 species of management concern in southwestern Wyoming. Report prepared for Advanced Resources International by the Wyoming Natural Diversity Database, University of Wyoming, Laramie. Available from http://uwadmnweb.uwyo.edu/WYNDD/Reports/pdf_beauvais/sw_wyo_predictive_rm_04.pdf

Drake JM, Bossenbroek JM (2004) The potential distribution of zebra mussels in the United States. Bioscience 54:931–941

Dunn DF (1981) The clownfish sea anemones: Stichodactylidae (Coelenterata: Actiniaria) and other sea anemones symbiotic with pomacentrid fishes. Trans Am Phil Soc 71(1):1–115

Fautin DG, Allen GR (1992) Field guide to anemonefishes and their host sea anemones. Western Australian Museum, Perth. Available from www.nhm.ku.edu/inverts/ebooks/intro.html

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49

Graybeal A (1998) Is it better to add taxa or characters to a difficult phylogenetics problem? Syst Biol 47:9–17

MacArthur RH (1972) Geographical ecology: patterns in the distribution of species. Harper & Row, New York

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence–absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Peterson AT (2001) Predicting species' geographic distributions based on ecological niche modeling. Condor 103:599–605

Peterson AT, Ball LG, Cohoon KP (2002) Predicting distributions of Mexican birds using ecological niche modeling methods. Ibis (online) 144:E27–E32

Raxworthy CJ, Martinez-Meyer E, Horning N, Nussbaum RA, Schneider GE, Ortega-Huerta MA, Peterson AT (2003) Predicting distributions of known and unknown reptile species in Madagascar. Nature 426:837–841

Semmens BX, Buhle ER, Salomon AK, Pattengill-Semmens CV (2004) A hotspot of non-native marine fishes: evidence for the aquarium trade as an invasion pathway. Mar Ecol Prog Ser 266:239–244

Shick JM (1991) A functional biology of sea anemones. Chapman & Hall, London

Soberón J, Peterson AT (2004) Biodiversity informatics: managing and applying primary biodiversity data. Phil Trans R Soc Lond Ser B Biol Sci 359:689–698

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148:1–13

Wiley EO, McNyset KM, Peterson AT, Robins CR, Stewart AM (2003) Niche modeling and geographic range predictions in the marine environment using a machine-learning algorithm. Oceanography 16:120–127

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: December 27, 2004; Accepted: December 12, 2005
Proofs received from author(s): June 14, 2006

Vol. 316: 285–310, 2006

MARINE ECOLOGY PROGRESS SERIES
Mar Ecol Prog Ser

Published July 3

Mapping world-wide distributions of marine mammal species using a relative environmental suitability (RES) model

K. Kaschner[1, 2, 3,*], R. Watson[1], A. W. Trites[2], D. Pauly[1]

[1]Sea Around Us Project, Fisheries Centre, University of British Columbia, 2259 Lower Mall, Vancouver, British Columbia V6T 1Z4, Canada

[2]Marine Mammal Research Unit, Fisheries Centre, University of British Columbia, Hut B-3, 6248 Biological Sciences Road, Vancouver, British Columbia V6T 1Z4, Canada

[3]Forschungs- und Technologiezentrum Westküste, Hafentörn, 25761 Büsum, Germany

ABSTRACT: The lack of comprehensive sighting data sets precludes the application of standard habitat suitability modeling approaches to predict distributions of the majority of marine mammal species on very large scales. As an alternative, we developed an ecological niche model to map global distributions of 115 cetacean and pinniped species living in the marine environment using more readily available expert knowledge about habitat usage. We started by assigning each species to broad-scale niche categories with respect to depth, sea-surface temperature, and ice edge association based on synopses of published information. Within a global information system framework and a global grid of 0.5° latitude/longitude cell dimensions, we then generated an index of the relative environmental suitability (RES) of each cell for a given species by relating known habitat usage to local environmental conditions. RES predictions closely matched published maximum ranges for most species, thus representing useful, more objective alternatives to existing sketched distributional outlines. In addition, raster-based predictions provided detailed information about heterogeneous patterns of potentially suitable habitat for species throughout their range. We tested RES model outputs for 11 species (northern fur seal, harbor porpoise, sperm whale, killer whale, hourglass dolphin, fin whale, humpback whale, blue whale, Antarctic minke, and dwarf minke whales) from a broad taxonomic and geographic range, using data from dedicated surveys. Observed encounter rates and species-specific predicted environmental suitability were significantly and positively correlated for all but 1 species. In comparison, encounter rates were correlated with <1% of 1000 simulated random data sets for all but 2 species. Mapping of large-scale marine mammal distributions using this environmental envelope model is helpful for evaluating current assumptions and knowledge about species' occurrences, especially for data-poor species. Moreover, RES modeling can help to focus research efforts on smaller geographic scales and usefully supplement other, statistical, habitat suitability models.

KEY WORDS: Habitat suitability modeling · Marine mammals · Global · GIS · Relative environmental suitability · Niche model · Distribution

INTRODUCTION

A number of marine mammal species are currently threatened by a variety of anthropogenic factors, ranging from bycatch and ship-strikes to pollution, global warming, and potential food competition (Perrin et al. 2002). The development and implementation of effective conservation measures require, however, detailed knowledge about the geographic occurrence of a species. In recent years, advances in geographic information systems (GIS) and computational power have allowed the development and application of habitat

[*]Email: kaschner@zoology.ubc.ca

suitability models to quantitatively delineate maximum range extents and predict species' distributions. Standard models rely on available occurrence records to investigate the relationships between observed species' presence and the underlying environmental parameters that — either directly or indirectly — determine a species' distribution in a known area and use this information to predict the probability of a species' occurrence in other areas (Guisan & Zimmermann 2000).

Habitat suitability models have been widely applied in terrestrial systems and for a wide range of land-based species (Peterson & Navarro-Sigüenza 1999, Zaniewski et al. 2002, Store & Jokimäki 2003). There are, however, comparatively few attempts to use such models to map species' distributions in the marine environment (Huettmann & Diamond 2001, Yen et al. 2004, Guinotte et al. 2006 in this Theme Section). This is particularly true for marine mammals, partly because the collection of species' occurrence data is hampered by the elusiveness and mobility of these animals. In addition, designated and costly surveys usually cover only a small fraction of a species' range (e.g. Kasamatsu et al. 2000, Hammond et al. 2002, Waring et al. 2002), due to the vastness of the marine environment and the panglobal distributions of many species. Thus, these surveys often yield little more than a snapshot, both in time and space, of a given species' occurrence. The comparatively low densities of many marine mammal species further contribute to the difficulties in distinguishing between insufficient effort to detect a species in a given area and its actual absence. On the other hand, a concentration of sightings may only reflect the concentration of effort rather than a concentration of occurrence (Kenney & Winn 1986).

There are on-going efforts — conducted, for example, as part of the OBIS initiative (Ocean Biogeographic Information System) — to compile existing marine mammal occurrence records, to allow for large-scale quantitative analyses of species distributions using habitat suitability modeling. For many species, however, there have been <12 known or published sightings to date. Actual point data sets, which generally cover only a fraction of known range extents, are available or readily accessible for <50 % of all marine mammal species through the OBIS-SEAMAP portal (http://seamap.env.duke.edu/), the currently most comprehensive data repository for marine mammal sightings.

As a consequence of this data paucity, marine mammal occurrence has been modeled for only a handful of species and only in relatively small areas. Most existing studies have employed so-called presence–absence statistical models, such as general linear models (GLMs) or general additive models (GAMs) (Moses & Finn 1997, Hedley et al. 1999, Gregr & Trites 2001, Hamazaki 2002). These model types require data collected during line-transect surveys that systematically document species' presences and absences to predict varying species' densities or probabilities of occurrence (Hamazaki 2002, Hedley & Buckland 2004). However, predictions from presence–absence type models are affected by species' prevalence (Manel et al. 2001). For marine mammals, however, densities and/or detectability tend to be very low. More importantly, representative survey coverage of entire range extents has currently been achieved for an estimated 2 % of all species. This precludes the application of presence–absence modeling techniques to predict occurrence on larger scales for the vast majority of all cetaceans and pinnipeds.

Ecological niche models such as GARP (Genetic Algorithm for Rule Set Production; Stockwell & Noble 1992) and ecological niche factor analysis (ENFA) (Hirzel et al. 2002) represent alternative approaches which — due to their more mechanistic nature — can reduce the amount of data needed, since they do not require absence data and may therefore use so-called opportunistic data sets. These presence-only models have found widespread application in terrestrial systems (Peterson et al. 2000, Peterson 2001, Engler et al. 2004), and, more recently, attempts have been made to use such models to predict distributions of some rarer marine mammal species (Compton 2004, MacLeod 2005). However, for most species, there are fewer occurrence records readily available than required to generate accurate predictions (e.g. 50 to 100 representative occurrence records in the case of GARP; Stockwell & Peterson 2002). Moreover, these niche models assume that data sets represent an unbiased sample of the available habitat (Hirzel et al. 2002), which makes them sensitive to the skewed distribution of effort prevalent in most opportunistically collected marine mammal data sets (see below).

In conclusion, the current shortage of point data sets has prevented applying standard empirical habitat suitability models to predict patterns of occurrences or maximum range extents on larger scales. Similarly, this lack of data has prohibited the prediction of occurrence patterns for the lesser-known marine mammal species in more inaccessible or understudied regions of the world's oceans — and will likely continue to do so in the foreseeable future. As a consequence, marine mammal distributional ranges published to date mainly consist of hand-drawn maps outlining the proposed maximum area of a species' occurrence based on the professional judgment of experts and synopses of qualitative information (e.g. Ridgway & Harrison 1981a,b, 1985, 1989, 1994, 1999, Perrin et al. 2002). Frequently, there is considerable variation amongst the range extents proposed by different authors for the same species (Jefferson et al. 1993, Reijnders et al. 1993). In addition, these maps are often supplemented by relatively large regions covered by question marks, indicating areas of unknown, but likely, occurrence. As an alternative, some authors have summarized available raw point data in the form of documented stranding or sighting locations on maps (e.g. Perrin et al. 1994, Jefferson & Schiro 1997, Ballance & Pitman 1998), thus leaving it to the readers to infer possible species' distributions. All of these approaches are greatly confounded by uncertainty in the degree of interpolation applied to the occurrence data (Gaston 1994), and none delineates species' distributions based on an explicit algorithm that captures patterns of species' occurrences using a rule-based approach or statistical models, as recommended by Gaston (1994).

Although we currently lack the comprehensive point data sets to remedy this situation using standard habitat suitability modeling techniques, we nevertheless already know quite a bit about the general habitat usage of most marine mammal species, available in the form of qualitative descriptions, mapped outlines, geographically fragmented quantitative observations, and large-scale historical catch data sets. Existing knowledge about species' occurrence is likely biased—given the high concentration of survey efforts in shelf waters of the northern hemisphere—and the lack of statistical investigations on resource selection does not allow definitive conclusions about habitat preferences for most species (Johnson 1980, Manly et al. 2002). However, the synthesis of available knowledge about species' occurrences, collected from wide range of sources, time periods, and geographic regions, may approximate a representative sampling scheme in terms of the investigation of habitat usage on very large scales—at least until sufficient point data sets become available for more rigorous analyses. In the meantime, we propose that expert knowledge may represent an alternative and underutilized resource that can form the basis for the development of other types of habitat suitability models, such as rule-based environmental envelope models. Envelope models and techniques relying on formalized expert opinion have frequently been used in the past to predict large-scale terrestrial plant distributions (e.g. Shao & Halpin 1995, Guisan & Zimmermann 2000, Skov & Svenning 2004), but have not yet been applied to describe marine mammal range extents.

The objective of this study was to develop a generic quantitative approach to predict the average annual geographical ranges of all marine mammal species within a single conceptual framework using basic descriptive data that were available for (almost) all species. We also wanted to gain insight into the potential relative environmental suitability (RES) of a given area for a species throughout this range. Since comprehensive point data sets are currently non-existent or non-accessible for the vast majority of marine mammal species, we sought to generate our predictions based on the synthesis of existing and often general qualitative observations about the spatial and temporal relationships between basic environmental conditions and a given species' presence. The maps we produced represent a visualization of existing knowledge about a species' habitat usage, processed in a standardized manner within a GIS framework and related to local environmental conditions. Thus, our results can be viewed as hypotheses about potentially suitable habitat or main aspects of a species' fundamental ecological niche, as defined by Hutchinson (1957). We tested and evaluated our model predictions and assumptions using available marine mammal sightings and catch data from different regions and time periods to establish the extent to which this approach may be able to capture actual patterns of species' occurrence. Finally, we explored the merits and limitations of the model as a useful supplement to existing habitat suitability modeling approaches.

MATERIALS AND METHODS

Model structure, definitions, scope, and resolution. We derived the geographic ranges for 115 marine mammal species and predicted the RES for each of them throughout this range based on the available information about species-specific habitat usage. We defined geographic range as the maximum area between the known outer-most limits of a species' regular or periodic occurrence. While this definition is inclusive of all areas covered during annual migrations, dispersal of juveniles etc., it specifically excludes extralimital sightings, which are sometimes difficult to distinguish from the core range (Gaston 1994). Adhering to the plea of Hall et al. (1997) for the use of clear definitions and standard terminology, we chose the term 'relative environmental suitability' rather than 'habitat suitability' to describe model outputs, to distinguish our predictions, which often corresponded more closely to a species' fundamental niche, from the actual probabilities of occurrence generated by other habitat suitability models (Hirzel et al. 2002).

General patterns of occurrence of larger, long-living animals, such as marine mammals, are unlikely to be affected by environmental heterogeneity over small temporal and spatial scales (Turner et al. 1995, Jaquet 1996). This may be especially true for species living in the marine environment, as pelagic systems show greater continuity in environmental conditions over evolutionary time than terrestrial environments (Platt & Sathyendranath 1992). We chose a global geographic scope to accommodate the wide-ranging

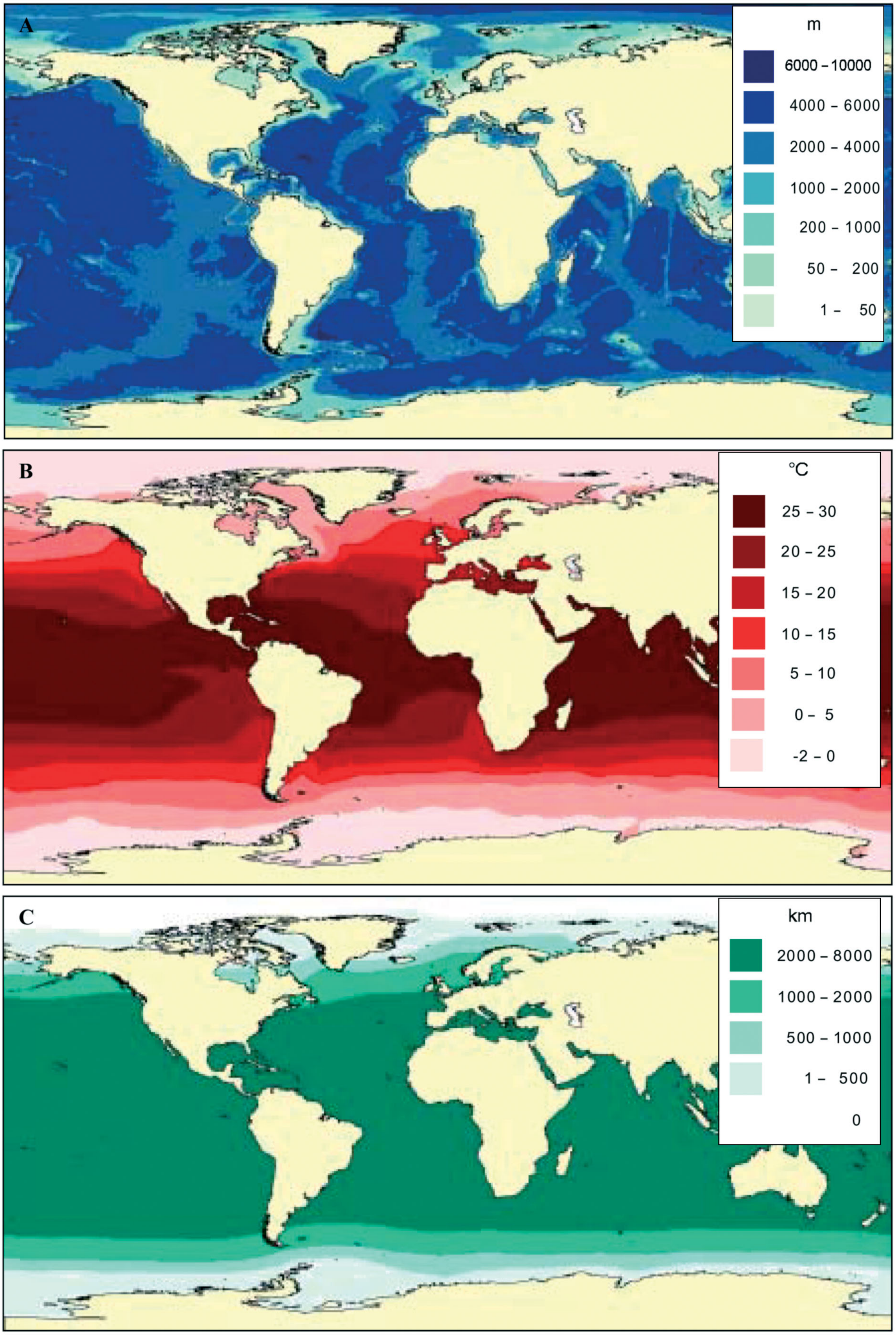

A
m
6000 – 10000
4000 – 6000
2000 – 4000
1000 – 2000
200 – 1000
50 – 200
1 – 50
B
°C
25 – 30
20 – 25
15 – 20
10 – 15
5 – 10
0 – 5
-2 – 0
C
km
2000 – 8000
1000 – 2000
500 – 1000
1 – 500
0

Fig. 1. Distribution of model predictors: (A) bathymetry (in m); (B) annual average sea-surface temperature (SST, in °C), and (C) mean annual distance to the ice edge (in km)

annual movements and cosmopolitan occurrence of numerous marine mammal species. Similarly, we used long-term averages of temporally varying environmental parameters to minimize the impacts of interannual variation. The model's spatial grid resolution of 0.5° latitude by 0.5° longitude represents a widespread standard for global models.

Independent variables. The lack of point data used for model input precluded the application of standard techniques to determine which environmental predictors might be best suited to predict species' occurrence. Instead, selection of environmental proxies that served as independent variables in our model was based on the existing knowledge about their relative importance to — indirectly — determine species occurrence for many marine mammals. Furthermore, predictors were chosen based on the availability of data at appropriate scales, including the availability of matching habitat usage information that was obtainable for all or at least the majority of all species. All environmental data were interpolated and rasterized using a custom GIS software package (SimMap 3.1 developed by R. Watson & N. Hall) and stored as attributes of individual grid cells in the global raster (Watson et al. 2004) (Fig. 1A–C).

Bottom depth: Strong correlations between bathymetry and patterns of inter- or intraspecific species' occurrences have been noted for many species of cetaceans and pinnipeds in different regions and ocean basins (Payne & Heinemann 1993, Moore et al. 2002, Baumgartner et al. 2001, Hamazaki 2002), making seafloor elevation an ideal candidate as an environmental proxy for a generic habitat suitability model. Bathymetric data were taken from the ETOPO2 dataset available on the United States National Geophysical Data Center's 'Global Relief' CD (www.ngdc.noaa.gov/products/ngdc_products.html), which provides elevation in 2 min intervals for all points on earth (Fig. 1A).

Mean annual sea-surface temperature: In addition to non-dynamic parameters, such as bathymetry, marine mammal distributions are influenced by a host of variable environmental factors, such as sea-surface temperature (SST). Changes in SST may be indicative of oceanographic processes that ultimately determine marine mammal occurrence across a number of different temporal scales (Au & Perryman 1985), and significant correlations of marine mammal species with SST have been demonstrated in different areas and for a variety of different species (e.g. Davis et al. 1998, Baumgartner et al. 2001, Hamazaki 2002). Surface temperature may not be a good predictor for all marine mammals, given the substantial foraging depths of some species (Jaquet 1996). However, we nevertheless chose to use SST as a proxy, because of the general availability of observations of surface climatic conditions or quantitative measurements associated with marine mammal occurrences. Global annual SST data, averaged over the past 50 yr, were extracted from the NOAA World Ocean Atlas 1998 CD (NOAA/NODC 1998) (Fig. 1B).

Mean annual distance to ice edge: The shifting edge of the pack ice is a highly productive zone (Brierley et al. 2002, Hewitt & Lipsky 2002) and represents important feeding grounds for many species of marine mammals (Murase et al. 2002). A number of studies have shown that sea ice concentration and ice cover, in combination with depth, play a key role in ecological niche partitioning for many species (Ribic et al. 1991, Moore & DeMaster 1997). We included the distance to the ice edge as an additional predictor in our model, as the distribution of species in the polar zones may not be fully captured using only SST. Although ice extent is strongly spatially correlated with SST, the actual edge of the sea ice does not directly coincide with any single isotherm throughout the year (Fig. 1B,C). Moreover, the ability of different marine mammal species to venture into pack-ice varies substantially. Spatial information about the average monthly ice extent (1979 to 1999) — defined by the border of minimum 50% sea ice coverage — was obtained from the United States National Snow & Ice Data Center (NSIDC) website (http://nsidc.org/data/smmr_ssmi_ancillary/trends.html#gis). We smoothed the ice edge border to correct some obvious misclassification and/or re-projection errors. After rasterizing the ice extent data, we calculated monthly distances from the nearest ice edge cell for each cell in the raster and computed annual average distances based on these monthly distances (Fig. 1C).

Distance to land: Some pinniped species — specifically the eared seals (otariids) — appear to be restricted to areas fairly close to their terrestrial resting sites, i.e. haulouts and rookeries (Costa 1991, Boyd 1998). The maximum distances away from these land sites are determined by a combination of species-specific life-history and physiological factors, such as the maximum nursing intervals based on the ability of pups to fast (Bonner 1984) and maximum swimming speed of adults (Ponganis et al. 1992). Global data sets identifying pinniped rookery sites do not exist. However, distance from landmasses in general was deemed to be an appropriate proxy in the context of this model and served as an additional predictor to more realistically model the distribution of some of the pinniped species (Appendix 2 in Kaschner 2004). For each cell, distance to land, defined as the nearest cell containing a part of coastline, was calculated in the same manner as distance to the ice edge.

Dependent variables. ***Marine mammal species:*** Our model encompassed 115 species of marine mammals that live predominantly in the marine environment (Table 1, present paper, and Appendix 1 in Kaschner 2004). We did not consider exclusively freshwater cetaceans or pinnipeds, nor the marine sirenians, sea otters, or the polar bear. Taxonomically, we largely followed Rice (1998), except for right whales, for which we recognized 3 separate species (Rosenbaum et al. 2000, Bannister et al. 2001). In addition, we included a recently described additional species, Perrin's beaked whale *Mesoplodon perrini* (Dalebout et al. 2002).

Definition of habitat usage or niche categories: Habitat usage categories were defined to represent broad predictor ranges, which roughly describe real marine physical/ecological niches inhabited by different marine mammal species. Niche categories effectively represent species response curves in relation to available habitat. Normally such response curves are derived empirically based on the statistical analysis of animal occurrences in relation to direct or indirect ecological gradients (Guisan & Zimmermann 2000, Manly et al. 2002). However, again, for the vast majority of marine mammal species the possible shape of such relationships remains to be investigated, and in the few existing studies only a sub-set of the available habitat has been covered (e.g. Cañadas et al. 2003). The more mechanistic nature of our model and the non-point type input data used precluded the derivation of empirical generic relationships within the context of this study. We therefore assumed a trapezoidal response curve (Fig. 2). We selected this shape as the most broadly appropriate option to model annual average distributions, as it represents a compromise between the likely unimodal response curves for species with fairly restricted ranges and the probably more bi-modal shape for species undertaking substantial migrations. The selected shape meant that the relative environmental suitability was assumed to be uniformly highest throughout a species' preferred or mostly used parameter range ($\mathrm{Min_P}$ to $\mathrm{Max_P}$ in Fig. 2). Beyond this range, we assumed that suitability would generally decrease linearly towards the minimum or maximum thresholds for a species ($\mathrm{Min_A}$ or $\mathrm{Max_A}$ in Fig. 2). Suitability was set to zero outside the absolute minimum or maximum values.

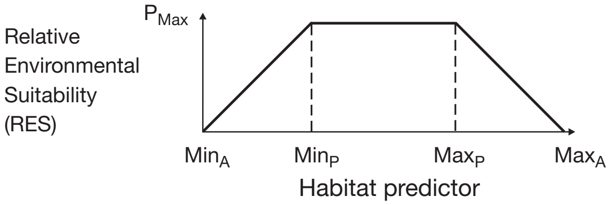


Fig. 2. Trapezoidal species' response curve describing the niche categories used in the RES model. $\mathrm{Min_A}$ and $\mathrm{Max_A}$ refer to absolute minimum and maximum predictor ranges, while $\mathrm{Min_P}$ and $\mathrm{Max_P}$ describe the 'preferred' range, in terms of habitat usage of a given species

While ecologically meaningful niches for bottom depth and association with ice extent are variable in

Table 1. Names, taxonomy, and general distributions of the 20 selected marine mammal species included in the relative environmental suitability (RES) model for which we show predictions (see Fig. 3) (for all other species see Kaschner 2004, her Appendix 1)

Common name	Scientific name	Suborder	Distribution
North Atlantic right whale	*Balaena glacialis*	Mysticeti	N Atlantic
Antarctic minke whale	*Balaenoptera bonaerensis*	Mysticeti	S hemisphere
Gray whale	*Eschrichtius robustus*	Mysticeti	N Pacific
Hourglass dolphin	*Lagenorhynchus cruciger*	Odontoceti	S hemisphere
Northern right whale dolphin	*Lissodelphis borealis*	Odontoceti	N Pacific
Irrawaddy dolphin	*Orcaella brevirostris*	Odontoceti	Indo-Pacific
Indian hump-backed dolphin	*Sousa plumbea*	Odontoceti	W Indian Ocean
Clymene dolphin	*Stenella clymene*	Odontoceti	Atlantic
Narwhal	*Monodon monoceros*	Odontoceti	Circumpolar, N hemisphere
S African & Australian fur seal	*Arctocephalus pusillus*	Pinnipedia	S Africa, S Australia
Guadalupe fur seal	*A. townsendi*	Pinnipedia	NE Pacific
New Zealand fur seal	*A. forsteri*	Pinnipedia	New Zealand, S Australia
Australian sea lion	*Neophoca cinerea*	Pinnipedia	S & SW Australia
South (American) sea lion	*Otaria flavescens*	Pinnipedia	S America
Galapagos sea lion	*Zalophus wollebaeki*	Pinnipedia	Galapagos Islands, E Pacific
Hooded seal	*Cystophora cristata*	Pinnipedia	N Atlantic
Ribbon seal	*Histriophoca fasciata*	Pinnipedia	N Pacific
Mediterranean monk seal	*Monachus monachus*	Pinnipedia	Mediterranean, NE Atlantic
Hawaiian monk seal	*M. schauinslandi*	Pinnipedia	Hawaii, NE Pacific
Ross seal	*Ommatophoca rossii*	Pinnipedia	Circumpolar, S hemisphere

width and were defined accordingly, SST categories were described by regular 5°C steps, based on the average intra-annual variation of 5 to 10°C in most areas of the world (Angel 1992). Quantitative definitions and corresponding qualitative descriptions of potential niches of the resulting 17 bottom depth ranges, 28 broad temperature ranges, and 12 ice edge association categories are shown in Table 2.

Table 2. Quantitative and qualitative definitions of habitat usage or niche categories (SST: sea-surface temperature; cont.: continental)

Environmental parameter	Minimum	Preferred		Maximum	Habitat category description
		minimum	maximum		
Depth usage zones (in m)	0	−1	−8000	−8000	All depths (uniform distribution)
	0	−1	−50	−200	Mainly estuarine to edge of cont. shelf
	0	−1	−50	−500	Mainly estuarine to beyond shelf break
	0	−10	−100	−1000	Mainly coastal−upper cont. shelf to upper cont. slope
	0	−10	−200	−2000	Mainly coastal−cont. shelf to end of cont. slope
	0	−10	−200	−6000	Mainly coastal−cont. shelf to deep waters
	0	−10	−1000	−6000	Mainly coastal−upper cont. slope to deep waters
	0	−10	−2000	−6000	Mainly coastal−cont. slope to deep waters
	0	−10	−2000	−8000	Mainly coastal−cont. slope to very deep waters
	0	−10	−4000	−8000	Mainly coastal−abyssal plains to very deep waters
	0	−200	−1000	−6000	Mainly upper cont. slope to deep waters
	0	−200	−2000	−6000	Mainly cont. slope to deep waters
	0	−200	−2000	−8000	Mainly cont. slope to very deep waters
	0	−200	−4000	−8000	Mainly cont. slope−abyssal plains to very deep waters
	0	−1000	−2000	−8000	Mainly lower cont. slope to very deep waters
	0	−1000	−4000	−8000	Mainly lower cont. slope−abyssal plains to very deep waters
	0	−2000	−6000	−8000	Mainly abyssal plains to very deep waters
Temperature usage zones (mean annual SST, in °C)	−2	−2	35	35	All temperatures (uniform distribution)
	−2	0	0	5	Polar only
	−2	0	5	10	Polar−subpolar
	−2	0	10	15	Polar−cold temperate
	−2	0	15	20	Polar−warm temperate
	−2	0	20	25	Polar−subtropical
	−2	0	25	30	Polar−tropical
	−2	0	30	35	Polar−full tropical
	0	5	5	10	Subpolar only
	0	5	10	15	Subpolar−cold temperate
	0	5	15	20	Subpolar−warm temperate
	0	5	20	25	Subpolar−subtropical
	0	5	25	30	Subpolar−tropical
	0	5	30	35	Subpolar−full tropical
	5	10	10	15	Cold temperate only
	5	10	15	20	Cold temperate−warm temperate
	5	10	20	25	Cold temperate−subtropcial
	5	10	25	30	Cold temperate−tropical
	5	10	30	35	Cold temperate−full tropical
	10	15	15	20	Warm temperate only
	10	15	20	25	Warm temperate−subtropical
	10	15	25	30	Warm temperate−tropical
	10	15	30	35	Warm temperate−full tropical
	15	20	20	25	Subtropical only
	15	20	25	30	Subtropical−tropical
	15	20	30	35	Subtropical−full tropical
	20	25	25	30	Tropical only
	20	25	30	35	Full tropical only
Ice edge usage zones (mean annual distance from ice edge, in km)	−1	0	8000	8000	No association with ice edge (uniform distribution)
	−1	0	500	2000	Mainly restricted to fast & deep pack-ice
	−1	0	500	8000	Mainly in fast & deep pack-ice, but also elsewhere
	0	1	500	2000	Mainly around edge of pack-ice
	0	1	500	8000	Mainly around edge of pack-ice, but also elsewhere
	0	1	2000	8000	Mainly in areas of max. ice extent, but also elsewhere
	0	1	8000	8000	Regularly but not preferably around edge of the pack-ice
	0	500	2000	8000	Mainly in areas of max. ice extent, but also elsewhere
	0	500	8000	8000	Regularly but not preferably in areas of max. ice extent
	500	1000	2000	8000	Mainly close to areas of max. ice extent
	500	1000	8000	8000	Regularly but not preferably close to max. ice extent
	1000	2000	8000	8000	No association with ice edge, nowhere near ice at any time of the year

Marine mammal habitat usages: We compiled published information about species-specific habitat usages with respect to their known association with the ice edge, as well as commonly inhabited bottom depth and SST ranges. Where appropriate, additional information about maximum likely distance from landmasses was also collected, based on information about maximum foraging trip lengths. Selected sources of information included >1000 primary and secondary references, all screened for relevant information on habitat use (compiled in Kaschner 2004, Appendix 2). Data extracted from these sources ranged from statistically significant results of quantitative investigations of correlations between species' occurrence and environmental predictors (e.g. Gregr & Trites 2001, Moore et al. 2002, Baumgartner et al. 2003, Cañadas et al. 2003), opportunistic observations (e.g. Carlström et al. 1997), maps of sightings or distribution outlines, to qualitative broad descriptions of prevalent occurrence such as 'oceanic, subtropical species' (e.g. Jefferson et al. 1993). A level of confidence was assigned to each record to reflect the origin, reliability, and detail of the data, with quantitative investigations of environmental factors and species' occurrence ranking highest and qualitative descriptions ranking lowest.

We assigned each species to niche categories for depth, temperature, and ice edge association (and in some cases distance to land) based on the most reliable information available (Table 3, present paper, and Kaschner 2004, Appendix 2). If the available information was inconclusive, or different conclusions could be drawn from the data, the species was assigned to multiple alternative niche categories representing different hypotheses. Distance from land preferences were used as an additional constraining factor for all species marked by an asterisk in Table 3 (present paper) and in Appendix 2 (Kaschner 2004). For a few species (<5), the general temperature categories were adjusted to reflect the extreme narrowness of their niche.

Area restrictions: On a global scale, contemporary distributions of marine mammals and other species are the result of their evolutionary history. Present occurrences and restrictions to certain areas therefore reflect a species center of origin and ability to disperse defined by its ecological requirements and competitors (LeDuc 2002, Martin & Reeves 2002). Information about a species' restriction to large ocean basins (i.e. North Atlantic or southern hemisphere), therefore, served as a rough first geographical constraint in the RES prediction model for each species to capture the results of this evolutionary process. The restriction to general ranges corresponds to the first-order selection of species in terms of habitat usage as described by Johnson (1980), and is implicitly incorporated in the sampling designs of many investigations of species' occurrence (Buckland et al. 1993).

If generated RES predictions did not reflect documented species' absences from certain areas, further geographical restrictions were imposed (Table 3, 'excluded areas'). It should be noted, however, that such restrictions were only imposed when known areas of non-occurrence were clearly definable, such as 'marginal' ocean basins (e.g. Red, Mediterranean, or Baltic Seas) or RES predictions showed signs of bi- or multimodality, meaning that areas of high suitability were separated by long stretches of less suitable habitat. We minimized introductions of such additional constraints so as not to impede the assessment of the ability of the RES model to describe, on its own, patterns of species' presence and absence.

Model algorithm — resource selection function. In our global raster, we generated an index of species-specific relative environmental suitability of each individual grid cell by scoring how well its physical attributes matched what is known about a species' habitat use. RES values ranged between 0 and 1 and represented the product of the suitability scores assigned to the individual attributes (bottom depth, SST, distance from the ice edge, and, in some cases, from land), which were calculated using the assumed trapezoidal response curves described above. A multiplicative approach was chosen to allow each predictor to serve as an effective 'knock-out' criterion (i.e. if a cell's average depth exceeded the absolute maximum of a species' absolute depth range, the overall RES should be zero, even if annual STT and distance to ice edge of the cell were within the species preferred or overall habitat range).

Multiple hypotheses about species distributions were generated using different combinations of predictor category settings if a species had been assigned to multiple, equally plausible, options of niche categories based on available data. The lack of test data sets for most species precluded the application of standard model evaluation techniques to determine the best model fit (Fielding & Bell 1997). Consequently, we selected the hypothesis considered to represent the best model fit through an iterative process and by qualitative comparison of outputs with all available information about the species' distribution and occurrence patterns within its range. Objective geographic ranges of species can then be determined based on some pre-defined threshold of predicted low or non-suitability of areas for a given species.

Model evaluation — species response curves and impact of effort biases. To assess the validity of using the RES model instead of available presence-only models, we investigated the degree to which available opportunistic data sets — for species with global or semiglobal distributions — may meet the basic assumption of existing niche models, i.e. unbiased effort coverage. The commercial whaling data is one of the largest opportunistic data sets of marine mammal occurrence, spanning almost 200 yr and approximating global coverage. Whaling operations did not adhere to any particular sampling schemes, and effort distributions were likely strongly biased. Nevertheless, it has been argued that such long-term catch data sets may still serve as good indicators of annual average species distribution and may thus provide some quantitative insight into general patterns of occurrence (Whitehead & Jaquet 1996, Gregr 2000). Consequently, whaling data would seem to be an obvious candidate for predicting distributions of marine mammal species with cosmopolitan or quasi-cosmopolitan range extents using existing presence-only modeling techniques. Using this data, we wanted to assess potential effort biases by comparing large-scale species response curves to environmental gradients derived from opportunistic and non-opportunistic data sets. In addition, we wanted to use the obtained response curves to evaluate the generic trapezoidal shape of our niche categories and how well habitat usage deduced from point data would correspond to the general current knowledge about such usages of specific species, as represented by the assigned niche category.

The opportunistically collected whaling data set contained commercial catches of member states of the International Whaling Commission (IWC) between 1800 and 2001 and was compiled by the Bureau of International Whaling Statistics (BIWS) and the Museum of Natural History, London, UK (IWC 2001a). We analyzed whaling data following an approach similar to that taken by Kasamatsu et al. (2000) and Cañadas et al. (2002) when investigating cetacean occurrence in relation to environmental gradients and generated species' response curves for 5 species with quasi-cosmopolitan distributions, including sperm whales *Physeter macrocephalus*, blue whales *Balaenoptera musculus*, fin whales *Balaenoptera physalus*, humpback whales *Megaptera novaeangliae*, and dwarf minke whales *B. acutorostrata*. The dwarf minke whale occurs to some extent sympatrically with its closely related sister species, the Antarctic minke whale *B. bonaerensis*. However, the 2 species are generally not distinguished in most data sets, and the analysis conducted therefore relates to a generic minke whale. As a first step, we assigned all catches recorded with accurate positions to the corresponding cell in our global raster, thus obtaining information about mean depth, SST, and distance to ice edge associated with each catch position. We then plotted frequency distributions of globally available habitat and the amount of habitat covered by whaling effort as the percent of total cells falling into each environmental stratum (defined to correspond to breakpoints in our niche categories) for depth, SST, and ice edge distance, to assess the extent to which whalers may have sampled a representative portion of the habitat available to species with global distributions.

To further assess potential effort biases, we generated histograms of catch 'presence' cells for individual species. These were based on the number of cells for which any catch of a specific species was reported within an environmental stratum and essentially represent visualizations of this species' response curve in relation to an environmental gradient. We then compared histograms based on catch 'presence' cells with both encounter rate distributions obtained from a non-opportunistic data set and catch distributions corrected for effort using an effort proxy developed during this study.

The non-opportunistic data set was collected during the IDCR/SOWER line-transect surveys, conducted annually over the past 25 yr in Antarctic waters and stored in the IWC-DESS database (IWC 2001b). Similar to the treatment of whaling data, we binned sighting records by raster cells, using only those records with sufficient spatial and taxonomic accuracy (i.e. sighting positions of reliably identified species were reported to, at least, the nearest half degree latitude or longitude). We then calculated species-specific encounter rates or SPUEs (sightings per unit of effort) across all years by computing total length of on-effort transects within each cell using available information about transect starting and end points. Finally, we plotted average SPUEs per environmental stratum to show species-specific response curves based on effort-corrected data.

To test if we could compensate for the absence of effort information in the opportunistic whaling data set, we derived a relative index of SPUE using a proportional sighting rate based on the fraction of total sightings in each cell that consisted of the specific species in question. We generated and compared proportional and standard encounter rates for dedicated IWC-IDCR survey data for a number of species. Both types of encounter rate were significantly and positively correlated for most species (e.g. $p < 0.0001$, Spearman's rho = 0.88 for minke whales). These results indicated that the developed effort proxy might indeed represent a good approximation of SPUE or CPUE (catch per unit effort) for data sets with missing effort information if multiple species were surveyed simultaneously. Based on the assumption that whalers would have caught any species of whale where and whenever they encountered it, we subsequently computed proportional catch rates for individual species for each cell using the whaling data set and were thus able to

Table 3. Habitat usage in terms of depth, mean annual SST, and distance to the edge of sea ice for selected marine mammal species. Superscripts denote the particular habitat type about which the reference provided information: ^{a}depth usage, ^{b}temperature usage, and ^{c}distance to edge of sea ice. For species marked by asterisk, distance from land was used as an additional constraining factor, limiting species to waters <500 km (*) from land (cont.: continental; estuar.: estuarine; reg.: regularly; pref.: preferably; assoc.: association; max.: maximum; Med: Mediterranean Sea; Black S.: Black Sea)

Common name	Depth range	Temperature range	Distance to ice edge range	General area minus (excluded areas)	Sources
North Atlantic right whale	Mainly coastal–continental shelf to deep waters	Subpolar–tropical	Mainly close to areas of max. ice extent	N Atlantic – (Black S., Med, Hudson Bay & Strait, Baltic)	Baumgartner et al. (2003)^{a}, Evans (1980)^{a}, Gaskin (1991)^{b}, Jefferson et al. (1993)^{c}, Kenney (2002)^{b}, Knowlton et al. (1992)^{a}, Mitchell et al. (1983)^{b}, Woodley & Gaskin (1996)^{a}
Antarctic minke whale	Mainly cont. slope to very deep waters	Polar–tropical	Mainly around edge of pack-ice, but also elsewhere	S hemisphere	Kasamatsu et al. (2000)^{a}, Murase et al. (2002)^{a,c}, Perrin & Brownell (2002)^{a,c}, Ribic et al. (1991)^{b}, Rice (1998)^{b,c}
Gray whale	Mainly estuar. to beyond shelf break	Subpolar–subtropical	Reg. but not pref. around edge of pack-ice	N Pacific	Deecke (2004)^{a,b}, Gardner & Chavez-Rosales (2000)^{b}, Jones & Swartz (2002)^{a,b,c}, Moore & DeMaster (1997)^{a,c}, Moore (2000)^{c}, Rugh et al. (1999)^{c}, Weller et al. (2002)^{a,b}
Hourglass dolphin	Mainly lower cont. slope–abyssal plains to very deep waters	Polar–warm temperate	Mainly in areas of max. ice extent, but also elsewhere	S hemisphere	Gaskin (1972)^{b}, Goodall (2002)^{a,b}, Goodall (1997)^{a,b,c}, Jefferson et al. (1993)^{a,c}, Kasamatsu et al. (1988)^{b}, Kasamatsu & Joyce (1995)^{c}
Northern right whale dolphin	Mainly lower cont. slope–abyssal plains to very deep waters	Subpolar–subtropical	No assoc. with ice edge, nowhere near ice	N Pacific – (Lat: <10° N) at any time of the year	Forney & Barlow (1998)^{a}, Jefferson & Newcomer (1993)^{a}, Jefferson et al. (1993)^{a}, (1994)^{c}, Rice (1998)^{c}, Smith et al. (1986)^{b}
Irrawaddy dolphin	Mainly estuar. to end of cont. shelf	Full-on tropical	No assoc. with ice edge, nowhere near ice at any time of the year	World – (Lon: >156° E & <80° E)	Arnold (2002)^{a,b}, Freeland & Bayliss (1989)^{a}, Mörzer Bruyns (1971)^{b}, Parra et al. (2002)^{a,b}, Rice (1998)^{c}, Stacey (1996)^{a,b}
Indian hump-backed dolphin	Mainly estuar. to end of cont. shelf	Subtropical–full tropical	No assoc. with ice edge, nowhere near ice at any time of the year	World – (Med., Black S. Lon >90° E & <14° E)	Findlay et al. (1992)^{a}, Jefferson et al. (1993)^{b}, Jefferson & Karczmarski (2001)^{a}, Karczmarski et al. (2000)^{a}, Rice (1998)^{c}, Ross (2002)^{a,b}
Clymene dolphin	Mainly cont. slope–abyssal plains to very deep waters	Full tropical only	No assoc. with ice edge, nowhere near ice at any time of the year	Atlantic – (Lon: >15° E & >70° W)	Davis et al. (1998)^{a,b}, Mullin et al. (1994)a^{a,b}, Perrin et al. (1981)^{a}, Rice (1998)^{c}
Narwhal	Mainly upper cont. slope to deep waters	Polar only	Mainly restricted to fast & deep pack-ice	N hemisphere	Dietz & Heide-Jørgensen (1995)^{a}, Heide-Jørgensen (2002)^{a,b}, Heide-Jørgensen et al. (2003)^{a}, Jefferson et al. (1993)^{b}, Martin et al. (1994)^{a}, Rice (1998)^{c}
Guadalupe fur seal*	Mainly lower cont. slope to very deep waters cont. slope	Warm temperate–tropical	No assoc. with ice edge, nowhere near ice at any time of the year	NE Pacific – (Lat: <10° N & Lon: >150° W)	Belcher & Lee (2002)^{b}, Lander et al. (2000)^{a}, Reijnders et al. (1993)^{b}, Rice (1998)^{c}

Table 3 (continued)

Common name	Depth range	Temperature range	Distance to ice edge range	General area minus (excluded areas)	Sources
S African & Australian fur seal*	Mainly coastal–upper cont. shelf to upper cont. slope	Warm temperate–subtropical	No assoc. with ice edge, nowhere near ice at any time of the year	S hemisphere – (Lon: >160° E & >20° W)	Arnould & Hindell (2001)[a], Reijnders et al. (1993)[b], Rice (1998)[c], Thomas & Schulein (1988)[a]
New Zealand fur seal*	Mainly coastal–cont. shelf to deep waters	Subpolar–warm temperate	Mainly close to areas of max. ice extent	S hemisphere – (Lon: >180°E & <150°E)	Bradshaw et al. (2002)[a], Jefferson et al. (1993)[b], Lalas & Bradshaw (2001)[a], Reijnders et al. (1993)[a], Rice (1998)[c]
Australian sea lion	Mainly coastal–upper cont. shelf to upper cont. slope	Warm temperate–subtropical at any time of the year	No assoc. with ice edge, nowhere near ice	S hemisphere – (Lon: >155°E & <75°E)	Costa (1991)[a], Gales et al. (1994)[b], Jefferson et al. (1993)[a], Ling (2002), Rice (1998)[c]
South (American) sea lion*	Mainly estuar. to end of cont. shelf	Polar–subtropical	Mainly close to areas of max. ice extent	S hemisphere – (Lat: >60°S & Lon: <40°W & >120°W)	Campagna et al. (2001)[a], Jefferson et al. (1993)[b], Reijnders et al. (1993)[b], Rice (1998)[c], Thompson et al. (1998)[a], Werner & Campagna (1995)[a]
Galapagos sea lion*	Mainly coast.–cont. shelf to deep waters	Full tropical only	No assoc. with ice edge, nowhere near ice at any time of the year	E Pacific – (Lat: >10°N & Lon: >100°W)	Dellinger & Trillmich (1999)[b], Heath (2002)[a], Jefferson et al. (1993)[a], Rice (1998)[c]
Hooded seal	Mainly lower cont. slope to very deep waters	Polar–cold temperate	Mainly around edge of pack-ice, but also elsewhere	N Atlantic	Folkow & Blix (1995)[a,c], Folkow et al. (1996)[a,c], Folkow & Blix (1999)[a], Kovacs & Lavigne (1986)[a,b,c], Reijnders et al. (1993)[b], Rice (1998)[c]
Ribbon seal	Mainly coast.–cont. slope to deep waters	Polar–subpolar	Mainly in areas of max. ice extent, but also elsewhere	N Pacific	Fedoseev (2002)[a,b], Jefferson et al. (1993)[a,b], Mizuno et al. (2002)[b], Reijnders et al. (1993)[a], Rice (1998)[c]
Hawaiian monk seal*	Mainly coast.–cont. shelf to deep waters	Subtropical–tropical	No assoc. with ice edge, nowhere near ice at any time of the year	NE Pacific – (Lat: <10°N & Lon: <140°W)	Gilmartin & Forcada (2002)[a], Parrish et al. (2000)[a], Parrish et al. (2002)[a], Reijnders et al. (1993)[b,c], Schmelzer (2000)[b]
Mediterranean monk seal	Mainly coastal–upper cont. shelf to upper cont. slope	Subtropical only	No assoc. with ice edge, nowhere near ice at any time of the year	N hemisphere – (Indian Ocean, Pacific, Lon: >20°W)	Duguy (1975)[a], Kenyon (1981)[a], Reijnders et al. (1993)[a,b,c]
Ross seal	Mainly coastal–cont. slope to deep waters	Polar only	Mainly restricted to fast & deep pack-ice	S hemisphere	Bengtson & Steward (1997)[a], Bester et al. (1995)[c], Jefferson et al. (1993)[b], Knox (1994)[b,c], Rice (1998)[c], Splettstoesser et al. (2000)[a], Thomas (2002)[c]

generate effort-corrected response curves of opportunistic whaling data.

Finally, we compared the 3 types of large-scale response curves for all 5 species and all predictors to assess impact of effort biases and to evaluate our choice of assigned niche categories and the generic trapezoidal niche category shape itself.

Model evaluation — RES model outputs. We evaluated the generated RES predictions by testing the extent to which these may describe the variations in actual species' occurrence for a number of marine mammal species found in different parts of the world's oceans using sightings and catch data collected during dedicated surveys. Species for which we tested predictions were harbor porpoises *Phocoena phocoena*, northern fur seals *Callorhinus ursinus*, killer whales *Orcinus orca*, hourglass dolphins *Lagenorhynchus cruciger*, southern bottlenose whales *Hyperoodon planifrons*, sperm whales, blue whales, fin whales, humpback whales, dwarf minke whales, and Antarctic minke whales. We selected species to cover a wide taxonomic, geographic, and ecological range to test the robustness of the generic RES approach. In addition, we chose test data sets that varied widely in geographic and temporal scope to assess at which temporal or spatial scale RES predictions may prove to be insufficient in capturing patterns of species' occurrences. To minimize risks of circularity, we tried to ascertain that test data had not been used to contribute directly or indirectly towards any of the studies or species reviews used to select input parameter settings. Test data sets included: (1) the SCANS (small cetaceans in the European Atlantic and North Sea) data collected during a dedicated line-transect survey in the North Sea and adjacent waters in the summer of 1994 (Hammond et al. 2002), (2) a long-term catch/sighting data set of northern fur seals collected during annual dedicated sampling surveys in the northeastern Pacific that were conducted in collaboration by the United States and Canadian federal fisheries agencies (Department of Fisheries and Oceans [DFO] — Arctic Unit & National Marine Fisheries Service [NFMS]) between 1958 and 1974, and (3) the long-term IWC-DESS data set described above (IWC 2001b) (Table 4).

Standard evaluation approaches for habitat suitability models based on confusion matrices are greatly impacted by difficulties to distinguish between true absences of species from an area and apparent absences due to detectability issues or insufficient sampling effort (Boyce et al. 2002). We therefore developed an approach similar one recommended by Boyce et al. (2002) to test predictions of presence-only models. Specifically, we compared the predicted gradient in RES scores across all cells covered by a survey with an observed gradient of relative usage by a given species in these cells, as described by the encounter rates of a species during the surveys. Again, species-specific encounter rates were obtained by binning records from each data set by raster cells, using only those records with sufficient spatial and taxonomic accuracy (i.e. catch or sighting positions of reliably identified species were reported to, at least, the nearest half degree latitude/longitude). For the reasons described above, we used the minke whale sightings in the IWC-DESS database to test the predictions for both the Antarctic minke whale and the dwarf minke whale.

Table 4. Sighting and catch data sets used for RES model testing (abbreviations for data sets and institutions see 'Model evaluation — RES model outputs')

	IWC-BIWS catch data	IWC-IDCR/SOWER survey data	SCANS survey data	Northern fur seal survey data
Agency/Source	IWC, UK, Bureau of Intern. Whaling Statistics, Norway & Natural History Mus. of London, UK	IWC member state collaboration	EU collaboration/ Sea Mammal Research Unit, UK	Arctic Unit, DFO, Canada & NMFS, US
Time period	1800–1999	1978–2001	June/July 1994	1958–1974
Survey area	World	Antarctica (south of 60°S)	greater North Sea	NE Pacific
Survey focal species	Large whales	Minke whales	Harbor porpoise	Northern fur seal
No. of marine mammal species reported	~20	~50	~5	1
No. of sighting/ catch records	~2 000 000	~35 000	1940	~18 000
Used for testing of	RES assumptions & model settings: minke, blue & humpback whale	RES results: Antarctic & dwarf minke, fin, blue & humpback whale, S. bottlenose whale, sperm & killer whale, hourglass dolphin	RES results: Harbor porpoise	RES: results: N. fur seal

Using only ship-based sightings, species-specific SPUEs were generated for the SCANS data set in the same fashion used for the IWC-DESS data. However, actual transect information was unavailable for the northern fur seal data set, although it contained absence records. Consequently, a proportional SPUE per raster cell was generated based on an approach similar to that applied to the IWC whaling data (i.e. we assumed that, on average, the total number of survey records [absence and presence] reported for 1 cell was representative of the effort spent surveying a cell).

For each test data set, we compared species-specific SPUEs with the corresponding RES model output for that species by averaging encounter rates over all cells covered by any effort that fell into a specific RES class. Using a bootstrap simulation routine, we generated 1000 random data sets, similar in terms of means, ranges, and distribution shapes to the predicted data set. We then used Spearman's non-parametric rank correlation test (Zar 1996, JMP 2000) to compare average observed encounter rates with corresponding RES classes based on model predictions and randomly generated data sets. To assess the performance of our model compared to random distributions, we obtained a simulated p-value by recording the number of times the relationship between random data sets and observed SPUEs was as strong as or stronger than that found between the observed encounter rates and our model predictions.

RESULTS

Relative environmental suitability predictions

Using available expert knowledge, RES modeling allows the prediction of potential distribution and habitat usage on very large-scales across a wide range of species in a standardized, quantitative manner. Model results represent specific, testable hypotheses about maximum range extents and typical occurrence patterns throughout a species' range averaged over the course of a whole year at any time from 1950 to 2000. Examples of RES predictions for 11 pinniped, 6 toothed, and 3 baleen whale species are shown in Fig. 3A–C. These examples were selected to demonstrate the applicability of the modeling approach over a wide geographic and taxonomic range of species (compare Table 1, present paper, with Kaschner 2004, her Appendix 1) and to illustrate the diversity of generated model outputs for species occupying different environmental niches. Where they existed, we included published outlines of maximum range extents (e.g. Jefferson et al. 1993, Reijnders et al. 1993) for comparison. RES predictions for all other species can be viewed on-line at www.seaaroundus.org/distribution/search.apx and are available in Kaschner (2004).

Generally, maximum extents of RES predictions for species closely matched published distributional outlines (Fig. 3). RES maps for many species also captured distinct areas of known non-occurrence well, without the need to introduce any geographic constraints. Examples of this are the predicted absence of hooded seals from Hudson Bay, the restriction of gray whales to the NE Bering Sea, and the non-occurrence of Irrawaddy dolphins in southern Australia.

RES modeling illustrates the degree of possible spatial niche partitioning that is already achievable based on the few basic environmental parameters. The complexity of the relationships between these parameters alone can lead to distinctly different patterns of suitable habitat for species with slightly different habitat usages, such as those demonstrated by the predictions for hooded seals (Fig. 3) and harp seal *Pagophilus groenlandica* in the North Atlantic (Kaschner 2004). Published maximum range extents of the 2 species, which are similar in terms of size and diets (Reijnders et al. 1993), suggest largely sympatric occurrences and a high degree of interspecific competition. However, small divergences in habitat usage of the 2 species (Table 3, present paper, and Kaschner 2004) resulted in predictions that suggest substantial spatial niche separation and highlight the importance of habitat preferences as a mechanism to reduce competition.

Model evaluation

Evaluation of species response curves and impacts of effort biases

Results from the analysis of whaling data highlighted the potential problems of using opportunistic data in presence-only models on very large scales in the marine environment. At the same time, results provided basic support for our selected niche category shape and the use of published information to assign species to niche categories.

Comparison of the distribution of catch 'presence' cells by environmental strata with globally available habitat indicated that even quasi-cosmopolitan and long-term opportunistic data sets such as the whaling data may not be a representative sub-sample of the habitat used by species with global range extents (Fig. 4A,B). Most existing presence-only models generate predictions based on the investigation of the frequency distribution of so-called presence cells in relation to environmental correlates. However, our analysis showed that simple species-specific catch 'presence' histograms that ignore the effects of hetero-

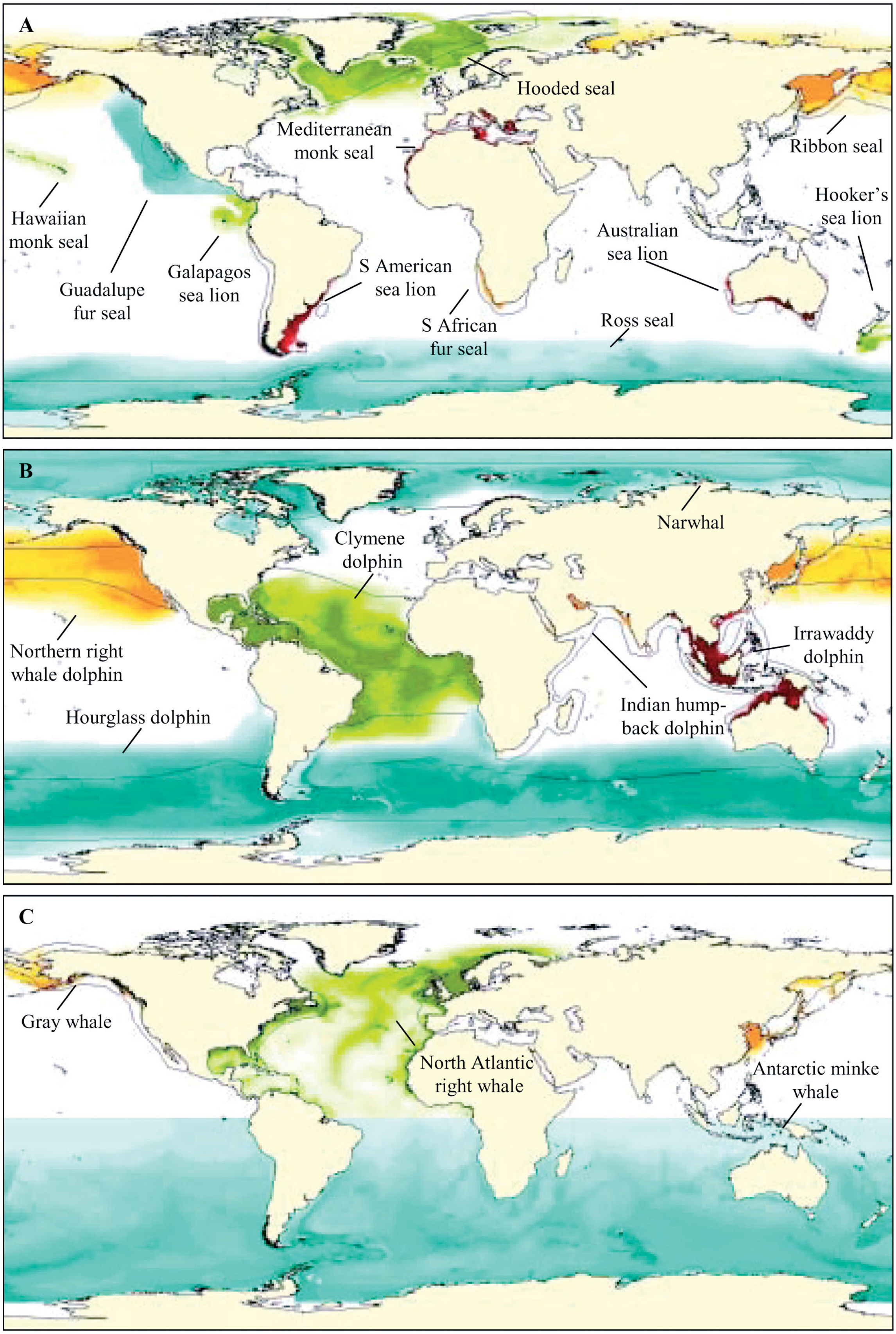

A
Hooded seal
Mediterranean monk seal
Ribbon seal
Hooker's sea lion
Hawaiian monk seal
Guadalupe fur seal
Galapagos sea lion
S American sea lion
Australian sea lion
S African fur seal
Ross seal
B
Narwhal
Clymene dolphin
Northern right whale dolphin
Irrawaddy dolphin
Hourglass dolphin
Indian hump-back dolphin
C
Gray whale
North Atlantic right whale
Antarctic minke whale

Fig. 3. Examples of RES model outputs: predicted RES (ranging from less suitable [light] to very suitable [dark]) based on habitat usage information for (A) 11 pinniped, (B) 6 odontocete and (C) 3 mysticete species. Outlines of proposed maximum range extent (Jefferson et al. 1993) are included for comparison. Note that, when viewed on a global scale, RES predictions for many coastal species are difficult to see in narrower shelf areas such as along the western coast of South America and eastern coast of Africa, and apparent absences from certain areas may just be artefacts of viewing scale. RES predictions of narwhal distribution in the Sea of Okhotsk are masked to some extent by those for the northern right whale dolphin. Similarly, predictions for New Zealand fur seals in Australia are masked by those for Australian sea lions. RES maps for all marine mammal species can be viewed on-line at www.seaaroundus.org/distribution/search.apx and are available in Kaschner (2004)

◄────

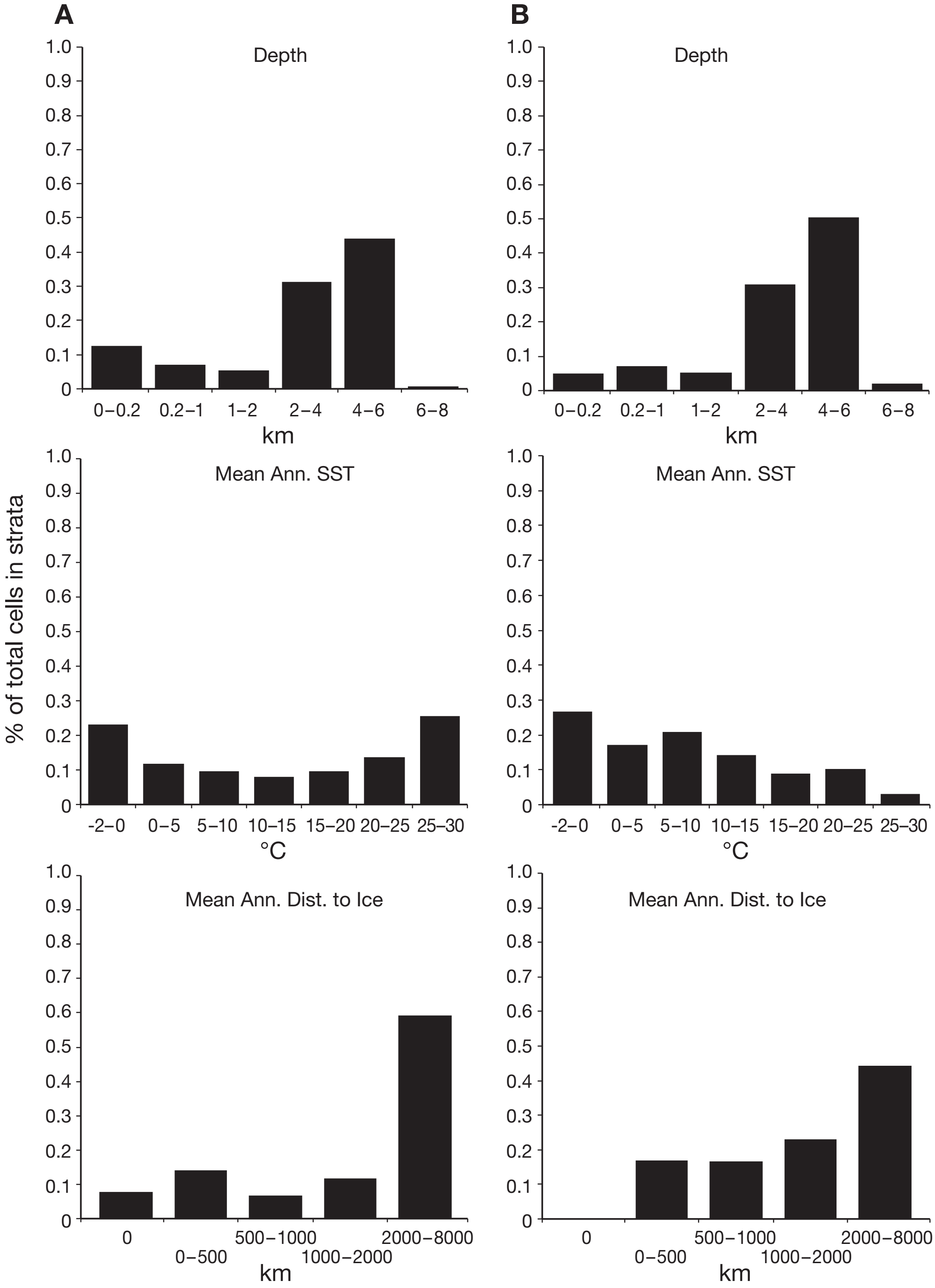


Fig. 4. Frequency distributions of: (A) globally available habitat and (B) amount of habitat covered by whaling effort as the percent of cells per available environmental stratum for depth, mean annual SST, and mean annual distance to ice edge

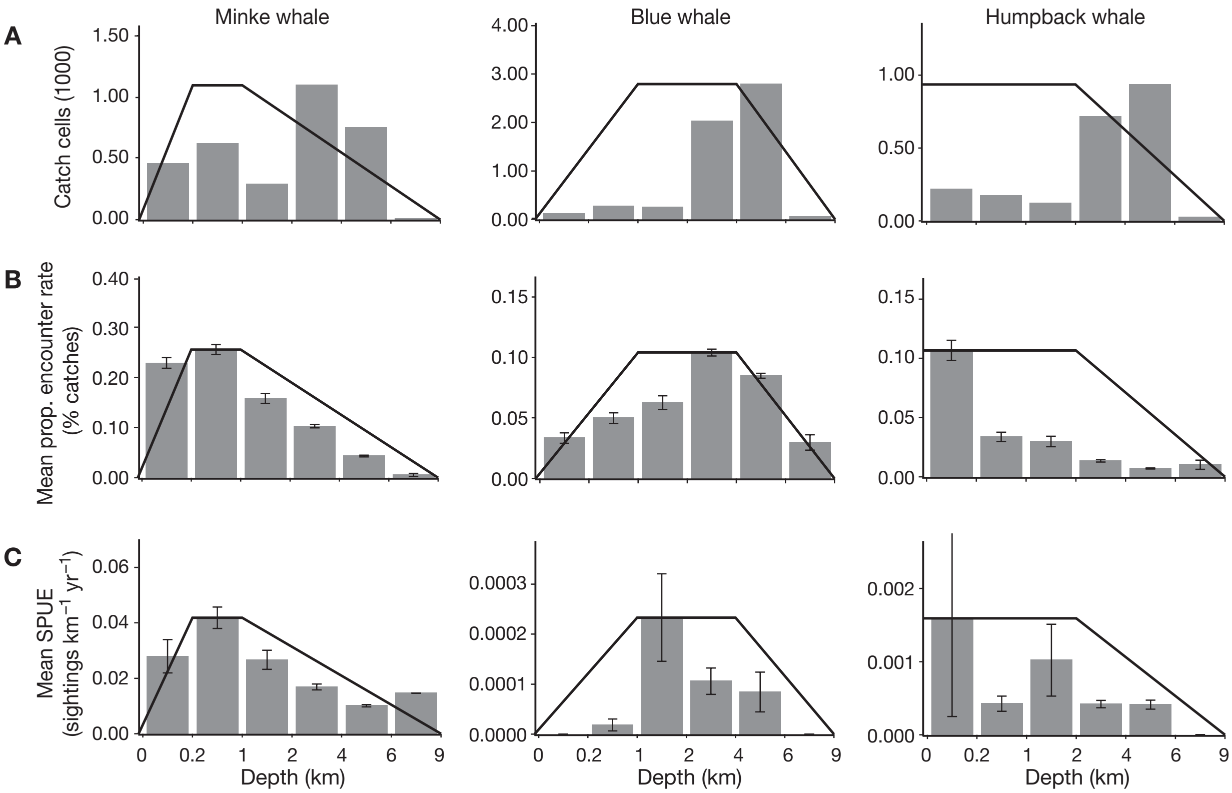


Fig. 5. Examples of depth usage of different globally occurring species using species' response bar plots. Plots were derived from IWC-BWIS whaling data and IWC-DESS dedicated survey data and illustrate the potential lack-of-effort biases introduced when using opportunistic point data sets for habitat suitability modeling. (A) Cumulative catch 'presence' cells per specified depth stratum (non-effort corrected), (B) same data after effort corrections using average proportional catch rates per stratum, (C) average sightings per unit effort (SPUE) per depth stratum obtained from dedicated surveys in Antarctic waters. Response plots based on effort-corrected opportunistic data closely resembled those derived from dedicated surveys. In contrast, relative depth usage based on catch presence cells alone would likely result in erroneous predictions of global species occurrence by presence-only habitat suitability models. Lines representing niche categories that species had been assigned to based on available published information (Table 3, present paper, and Appendix 2 in Kaschner 2004) were included to illustrate the extent to which response plots based on catch and sighting data supported our choice of niche category for each species. Note that response bar plots were scaled to touch top line for better visualization of niche category fit

geneously distributed sampling effort generally diverged substantially from bar plots of encounter rates obtained from dedicated survey data collected in the same area for all species investigated (see examples shown in Fig. 5A,C). In contrast, effort-corrected proportional catch rates by environmental strata closely resembled bar plots generated from dedicated survey data (Fig. 5B,C). Overall, all available information suggested that the trapezoidal shape of niche categories used in this model may be a reasonable approximation of marine mammal response curves for those species for which habitat usage could be investigated on larger scales.

In terms of depth ranges used, we generally observed a good fit between the niche categories we had assigned species to and the bar plots based on proportional catch rates and SPUEs, though not with those based on frequency distributions of catch 'presence' cells (Fig. 5). In contrast, with respect to temperature and distance to ice, we found great discrepancies between general current knowledge about the global habitat usage of many species and the respective species' habitat use that was suggested by all bar plots for these 2 predictors (not shown). These findings suggested that predictions of global, year-round distributions generated by standard presence-only modeling techniques and based on the whaling data alone might not reflect total distributional ranges of these species well.

Evaluation of RES predictions

RES modeling captured a significant amount of the variability in observed species' occurrences — corrected for effort—in all test cases (Table 5). Average species' encounter rates were positively correlated with predicted suitability of the environment for each species, except for

Table 5. Statistical results of model validation for different species including relevant information about test data sets to illustrate robustness of the RES model. Relationships between RES categories and average observed SPUEs were tested using Spearman's non-parametric rank correlation analysis. Simulated p-values represent the percentage of random data sets, generated using bootstrap simulation, that were more strongly correlated with observed data than RES predictions for given species (note that the analysis compared absolute strengths of correlations, i.e. in the case of the dwarf minke whale 0% of all random data sets were more strongly negatively correlated with the observed data). Note that generic 'minke whale' sightings were used to test RES predictions for the Antarctic minke and the dwarf minke whale

Common name	Survey area (1000 km^2)	Time period covered	No. of reported encounters	Results of rank correlation analysis of RES vs. SPUE		Comparison with random data sets
				rho	p	Simulated p-value
Northern fur seal	2	~20 yr	10 254	0.54	<0.0001	0
Harbor porpoise	0.7	~1 mo	1 265	0.59	<0.0001	0
Sperm whale	15	~20 yr	951	0.66	<0.0001	0
Killer whale	15	~20 yr	472	0.56	<0.0001	0.54
S. bottlenose whale	15	~20 yr	627	0.83	<0.0001	0
Hourglass dolphin	15	~20 yr	161	0.68	<0.0001	0
Antarctic minke whale	15	~20 yr	12 288	0.71	<0.0001	0
Dwarf minke whale	15	~20 yr	12 288	−0.77	<0.0001	0
Fin whale	15	~20 yr	163	0.53	<0.0001	0
Blue whale	15	~20 yr	72	0.48	<0.0001	0.268
Humpback whale	15	~20 yr	303	0.20	<0.05	0.006

the dwarf minke whale (Table 5). For this species, RES predictions were significantly but negatively correlated with the generic minke whale records in the IWC-IDCR data set. In contrast, <1% of the random data sets produced results that were more strongly correlated with observed encounter rates than the RES predictions in most cases (Table 5). Killer whales and blue whales were the only 2 species for which a higher percentage of random data sets showed an equally strong correlation with the observed SPUEs. Only for these 2 species chance cannot be excluded as a factor to explain the significance of the relationship detected between RES predictions and observed patterns of occurrence. Model predictions were fairly robust across a large range of temporal and spatial scales, as significant correlations were found even in the case of harbor porpoise using the comparatively small-scale and short-term SCANS data set.

DISCUSSION

RES predictions

Our model represents a new objective approach for mapping large-scale distributions of marine species using non-point data. Predictions represent the visualization of current expert knowledge about species occurrence with respect to some aspects of environmental heterogeneity that indirectly determine distribution boundaries and patterns of occurrence of species within these boundaries. RES model performance is convincing when compared to existing information about species' distributions, available in the form of descriptions of occurrences (see e.g. Rice 1998), or existing sketched outlines of distributional ranges (Jefferson et al. 1993). RES predictions are based on clearly defined assumptions and parameter settings and are thus reproducible and testable—unlike sketched distribution maps that may vary considerably between sources owing to differences in underlying assumptions or subjective and possibly arbitrary decisions made by the expert who drew them. In addition, by sacrificing 'detail for generality' (Levins 1966, Gaston 1994) and utilizing non-point data such as expert knowledge, the RES model can accommodate the frequently poor quality of available species' occurrence data that often precludes the use of other statistical habitat prediction approaches. Because our more process-orientated approach is based on information about a species' general occurrence in ecological space, like other niche models, it may be applied beyond existing survey ranges in geographic space (Hirzel et al. 2002). Thus, RES modeling represents a useful tool to investigate different hypotheses about large-scale distributions over a broad range of species, including those for which only few sighting records exist. In summary, the principle strength of the RES model lies in its greater objectivity in comparison to hand-drawn range extent and its generic applicability and its ability to utilize non-point data in comparison to statistical habitat suitability models.

In most cases, the predicted relative environmental suitability corresponded closely to the present ecological niche of a species. In other cases, predictions approximated a species' habitat, including its historical range extension prior to human-induced depletion. For some species, however, our results diverge substantially from known distributional ranges, suggesting that other factors may play a more important role in determining distributions. In general, RES predictions should be viewed as hypotheses about major aspects of a species' fundamental spatial niche.

RES predictions: limitations and biases

The predictions generated by our model are affected by various biases, operating at different levels. Some biases are inherent to the present implementation of our approach, such as the lack of consideration of other factors known to influence species' occurrence or the definition and shape of species response curves or the model algorithm. Other biases are directly associated with the data sets used for dependent and independent variables.

Other factors influencing species' occurrence

In most cases, the realized niche of a species is likely to be influenced by far more factors other than the 3 basic environmental parameters considered in our model, though the role these play will differ among species. Investigations of environmental correlates of species' occurrence have identified a host of other parameters, such as warm core rings for sperm whales (Jaquet & Whitehead 1996), zones of confluence of cyclone–anticyclone eddy pairs for a number of cetacean species (Griffin 1999, Davis et al. 2002), or the depth of the bottom mixed layer for North Atlantic right whales (Baumgartner et al. 2003). Consequently, it can be expected that the incorporation of factors such as these would lead to more heterogeneous patterns of species' occurrence than implied by our model results.

Dynamic ecological factors, such as intra- and interspecific competition and other behavioral interactions, also greatly influence the occurrence of species, especially on smaller geographic and temporal scales (Austin 2002). Such factors may considerably reduce niche overlap between different species as, for example, in the cases of Australian sea lions and New Zealand fur seals. These 2 species co-occur along the southern Australian coastline as implied by RES predictions (Fig. 4), but in reality occupy different niches within this region due to behavioral differences (Ling 1992).

RES modeling currently also ignores effects of seasonality and environmental regime shifts, as well as changes in habitat preferences or usage associated with different phases in the annual life cycle of a species. The lack of consideration of short-term and long-term temporal variation of environmental parameters will be most noticeable in areas with great inter-annual or seasonal fluctuations, such as for some areas along the east coast of the United States (Angel 1992, NOAA/NODC 1998) or during environmental regime shifts such as El Niño events. Likewise, discrepancies between known occurrences and RES predictions will be more pronounced for species undergoing extensive annual migrations or for those species with large increases or decreases in population size. Changes in habitat usages, well documented for many of the baleen whales (Kasuya & Miyashita 1997), often accompany the seasonal shifts from feeding to breeding grounds. Here, parameters other than those determining food availability may become important, such as predator avoidance (Corkeron & Connor 1999, Pitman et al. 2001). Similarly, it has been proposed that extreme fluctuations in population size and associated range depletions or expansions may result in changes in habitat usages over long temporal scales, especially in highly depleted, long-lived species such as the North Pacific right whale *Eubalaena japonicus* (Tynan et al. 2001).

Some of the most obvious discrepancies between RES predictions and known regional occurrences of species, however, may be explained by range depletions caused by past or present anthropogenic impacts, such as whaling, sealing, or fisheries bycatch. An example of the importance of this human-related factor is the stark contrast between the predicted distribution of the North Atlantic right whales (Fig. 3), and today's well-known absence of this species from northeastern Atlantic waters (Perry et al. 1999), due to exploitation by whalers in past centuries (Brownell et al. 1983).

Model algorithm biases

Observed discrepancies between RES predictions and known species' occurrences may also be due to biases inherent in the RES model algorithm and the assumptions about niche category shape and types, all of which are likely simplistic. A linear relationship between all 3 environmental parameters is improbable, as is the assumption that each of them will play an equally important role in influencing distributions across all species (as implied by our unweighted resource selection function). Likewise, the unimodal shape of niche categories — although found to be the most common type of functional responses in 1 terrestrial study (Oksanen & Minchin 2002) and to some extent supported by the investigation of large-scale species' response curves conducted here (Fig. 5) — is unlikely to adequately describe the presences of mammal species along environmental gradients in marine ecosystems in many cases. Although functional responses are probably strongly bimodal for some migratory species, the trapezoidal shape we used may, nevertheless, represent the most parsimonious and broadly applicable choice for predicting general annual average distributions. Investigations of marine mammal occurrence along environmental gradients in the past have been mainly restricted to relatively small scales, generally only encompassing a sub-set of the species' range (Baumgartner 1997, Kasamatsu et al. 2000, Cañadas et al. 2002). In the future, a meta-analytical investigation of large-scale response curves for some of the more data-rich marine mammal species would allow us to improve our current assumptions and is therefore regarded as a high priority.

Our quantitative definitions of niche categories currently ignore geographical differences in factors that determine niche boundaries or community transition zones. For instance, in comparison to other parts of the world, the edge of the continental shelf is typically much deeper (~500 m) in Antarctic waters, where the weight of the ice has caused the continental plate to sink (Knox 1994a). Consequently, the definition of 200 m bottom depth as a cut-off point for shelf-edge categories (Table 2) resulted in predicted absences of many species in some Antarctic regions where these species are known to occur regularly in high numbers (Hedley et al. 1999, IWC 2001b).

Biases of dependent and independent variables

The environmental parameters used as predictors in our model were affected by biases, which include direct measurement errors associated with the samples, and problems introduced through interpolation and rasterization processes (for detailed reviews of biases please refer to data providers, such as http://nsidc.org/data/smmr_ssmi_ancillary/trends.html#gis and NOAA/NODC 1998). Long-term averages of SST measurements will have been particularly affected by interpolation issues due to the temporally heterogeneous sampling effort over the past 50 yr (NOAA/NODC 1998). As a result, RES predictions may be biased towards time periods of higher sampling effort. Long-term ice edge data is affected by similar biases, but RES predictions were also influenced by the manual smoothing of ice edges, undertaken to eliminate nonsensical results in the computation of ice edge distances. In some cases, this smoothing resulted in predicted false absences or presences of species, such as the absence of harbor porpoise from the Baltic and Sea of Azov (Kaschner 2004). Furthermore, predictions were affected by the use of simple presence/absence ice data which did not allow the distinction between fast-ice (e.g. Weddell seals *Leptonychotes weddellii*; Kaschner 2004) and pack-ice species (e.g. Ross seal; Fig. 3). In the future, some ice data biases may be reduced by the use of more flexible sea ice concentration gradients instead of rigid presence/absence thresholds.

Unlike the independent variables, the information forming the basis for our dependent variables is less likely to be affected by interpolation issues, due to its mainly qualitative nature. Nevertheless, skewed effort distribution is likely to have had some influence on the current general perceptions about the habitat usage of many species.

Model evaluation

Evaluation of species response curves and impacts of effort biases

Investigation of the relationships between a species' occurrence and existing environmental gradients — which forms the basis of all habitat suitability models — requires adequate coverage of the habitat available to this species both in space and time (Manly et al. 2002). Comparison of the proportion of habitat covered by whaling operations with globally available habitat suggested that, even for very large opportunistic data sets such as the whaling data, sampling effort might not be equally representative of all habitat that is available to species with known cosmopolitan distributions. Though unbiased sampling effort is a key assumption also for presence-only models (Hirzel & Guisan 2002), predictions of terrestrial species' distributions generated by GARP, for instance, have been shown to be relatively insensitive to heterogeneously distributed effort (Peterson 2001, Stockwell & Peterson 2001). However, in comparison to terrestrial systems, insufficient coverage of available habitat due to spatially and temporally skewed effort is likely much more pronounced in the marine environment, where weather conditions and sheer distances restrict survey efforts mainly to the summer months and to areas relatively close to ports.

The importance of effort considerations was illustrated by the comparison of species' response curves to environmental gradients based on opportunistic data sets and those derived from effort-corrected data or available habitat usage information. Minke whales, for instance, are generally perceived to be closely associated with coastal and shelf waters (Jefferson et al. 1993) — a perception which is supported by statistical investigations of minke whale occurrences in relation to depth throughout the world (Sigurjónsson 1995, Kasamatsu et al. 2000, Hamazaki 2002, Moore et al. 2002) and is reflected by our choice of niche category.

However, this perception was greatly at odds with the depth distribution of minke whale catches in the whaling data, which — even if catch numbers were corrected for proportionally available habitat — suggested a predominant usage of much deeper waters for this species. The high number of minke whale catches reported in offshore areas might be explained by the concentration of whaling activities in deeper waters, where the larger whale species, such as blue, fin, and sperm whales that initially represented the primary targets of whalers, were predominantly known to occur (Perry et al. 1999). Minke whales did not become a target species until quite late in the whaling era, but were likely nevertheless caught on a regular basis if whalers happened upon them. The sheer amount of whaling effort in deeper waters thus masked this species' actual habitat usage if analyses were based on frequency of catch 'presence' cells alone.

In contrast, bar plots of effort-corrected catches were consistent with the general perception of depth usage of minke whales. The use of proportional encounter rates to investigate species' response curves might therefore help to compensate for some effort biases. In combination with results from other studies of cetacean response curves (e.g. Kasamatsu et al. 2000, Cañadas et al. 2002), bar plots of encounter rates based on both whaling data and dedicated surveys provided good support for the trapezoidal shape of niche categories used here.

In our analysis, we chose to ignore all temporal aspects of the data sets. The binning of catches across all years will have masked effects of the well-known serial depletion of the large whale species (Clark & Lamberson 1982, Perry et al. 1999) and the distortion likely introduced by any progressive spatial expansion of catch effort (Walters 2003). In view of these temporal biases and the very different time periods during which whaling data and the IWC-DESS survey data were collected, the similarity of encounter rate bar plots based on the 2 data sets was quite surprising. We propose that these findings provide indications that general usage of habitat by the species investigated here may have been quite consistent over the last century, despite the considerable fluctuating in population sizes.

The extent to which species' response curves from opportunistic data sets may be representative of habitat usage throughout a species' range appears to depend on the type of environmental predictor. The good fit of encounter rate bar plots and selected niche category in terms of bottom depth across almost all species indicated that whaling records indeed reflect the predominant perception of a species' global depth usage — if effort is taken into consideration. However, comparison of general current knowledge about global habitat usage in terms of temperature and ice distance — as represented by our selected niche categories for the different species — with bar plots for these 2 predictors suggested that catch data distributions were strongly seasonally biased. Whaling effort was concentrated in the polar waters of both hemispheres during summer months (IWC 2001a), thereby only covering parts of the distributions of most species targeted, namely their summer feeding grounds. While a species' depth preference is often consistent throughout its latitudinal range extent, temperature ranges and distance to ice edge will tend to vary depending on when and where throughout its range and annual life cycle an animal is captured or sighted. Thus, from the perspective of modeling highly migratory species with global distributions in the marine environment, reliance on available point data sets alone would likely result in a biased prediction, despite the potentially broad geographic coverage and large sample sizes of such data sets. In contrast, RES outputs may represent more balanced predictions of annual average distributions of cosmopolitan or quasi-cosmopolitan species, since we were able to supplement seasonally biased point data with additional sources of information about general occurrences during other times of the year when we assigned species to specific niche categories (Table 4).

In conclusion, our analyses of whaling data suggested that for habitat prediction on very large scales it may be difficult to find data sets that would allow the straightforward application of presence-only habitat suitability models. Nevertheless, a quantitative comparison of the quality of RES predictions for quasi-cosmopolitan marine species with those generated by other niche models using available opportunistic data sets is needed to allow a more rigorous investigation of the effects of skewed effort distributions on very large scales.

Evaluation of RES predictions

Statistical tests of RES model results indicated that our generic approach has some merit to adequately describe suitable habitat, as significant amounts of the variability in average species' occurrence were captured for all but 1 species tested (Table 5). In contrast, simulated random data sets rarely showed equally strong or stronger relationships with the observed data.

Several factors may explain the 2 cases in which random data sets often showed equally strong relationships with the observed data. For blue whales, the observed number of encounters was very low, possibly leading to the relatively weak correlation between predicted RES values and the test data set (Table 5). For killer whales, several different ecotypes or subspecies occupy distinctly different ecological niches in different parts of the world, including Antarctic waters (Pitman & Ensor 2003). To capture the preferred habitat of all subspecies, we selected very broad niche categories. Likewise, the IWC-IDCR data set does not distinguish between different subspecies, as these are difficult to identify in the wild. The very broad predictions and the mixed sightings pool of subspecies with different habitat usage may have contributed to the large proportion of random data sets that could explain the observed variation in the test data set equally well. Similarly to the mixed pool of killer whale sightings, the generic 'minke whale' observations in the test data set likely represent sightings of both the Antarctic minke whale and the dwarf minke whale — 2 species which appear to prefer slightly different habitats (IWC 2001b, Perrin & Brownell 2002, Matsuoka et al. 2003). Interestingly, RES predictions for the Antarctic minke whale were positively correlated with the generic sightings, while our predictions for the sister species showed an equally significant but negative correlation with these sightings. This suggests that either all minke whales encountered in the survey belonged to just 1 species, the Antarctic minke whale, or — and this is more likely — our model exaggerated the niche separation between the 2 species.

Independence of test data

The statistical testing of both our predictions and model assumptions are affected by a number of biases. First, given the broad nature of our niche categories and the type of information they were based on, we cannot be certain that the test data sets were indeed completely independent. Consequently, there is a risk of circularity, if the test data had somehow formed the basis of one of the broad 'expert knowledge' statements (such as 'coastal' and 'subtropical' species) that was fed into our model. However, the process of abstraction from point data to these general statements, in and of itself, would probably ensure a certain degree of data independence. Furthermore, we argue that—even if test data did serve as a basis of niche descriptions—testing the extent to which such broad statements may actually suffice to describe species' presences and absences when applied in a GIS modeling framework is a worthwhile exercise. Nevertheless, we tried to minimize potential circularity by excluding all references that were directly based on these data from our pool of input sources used to determine niche settings for the particular species tested (e.g. Kasamatsu et al. 2000, Hammond et al. 2002).

Comparison with other habitat suitability modeling approaches

The validation analysis indicated a remarkable robustness of RES predictions across a broad range of temporal and spatial scales and for a wide taxonomic range of species, suggesting that species' distributions and patterns of occurrence in the marine environment may be quantitatively described using surprisingly few basic parameters. Despite the apparent robustness of the RES modeling approach to perform well at different scales, care should be taken when interpreting model outputs.

It is highly unlikely that our more mechanistic model will be capable of predicting the real probability of species' occurrences in a specific place on a specific day or month of a given year. The RES model should therefore not be viewed as an alternative to empirical presence/absence type habitat prediction approaches that can and should be applied on smaller geographic scales to predict marine mammal occurrence when and where dedicated line-transect data sets are available. Similarly, the application of more sophisticated presence-only models, such as GARP or ENFA, may often be preferable at intermediate scales and when available data sets can be shown to represent a geographically and temporally unbiased subsample of the habitat available to a species. However, there is some indication — based on the analysis of whaling data — that effort biases might be more prominent in the marine environment than in terrestrial systems, thus potentially precluding the straightforward use of available opportunistic point data sets in presence-only models, though this remains to be investigated in more detail. In general, the quality of predictions generated by any model can only be as good as the available data, and more sophisticated models do not necessarily perform better than simpler approaches, especially if data quality is poor (Moisen & Frescino 2002). Consequently, RES modeling may be more suitable than other niche models on very large scales, where available data sets may not be representative of the species' actual occurrence or if point data are completely missing.

Future work and applications

In the future, RES modeling may serve as a useful tool to address both basic ecological questions as well as management and conservation-related issues in situations where the paucity of comprehensive point data sets — a situation commonly encountered in the marine environment — precludes the use of other more data-intensive habitat modeling approaches. Relying on more readily available types of data, such as expert knowledge, RES modeling will be particularly useful to study basic niche similarities and overlap between different marine species or groups of species on very large scales. Its application may also be a worthwhile first step in investigating scientific questions challenged by the paucity or complete lack of existing occurrence records, including historical distributions of heavily depleted species (e.g. gray whales in the North Atlantic; Mitchell & Mead 1977), calving grounds of endangered baleen whale species (yet unknown for species such as the North Pacific right whale; Gaskin 1991), or changes in species distributions due to environmental regime shifts or climate change (K. Kaschner unpubl. data).

Most importantly, however, the extent to which RES-generated hypotheses describe observed patterns in species' occurrence will allow more process-orientated questions to be asked about the role that other factors play in determining actual distributions. Similarly, the quantitative comparison of RES predictions with other niche models, such as GARP or ENFA, will help identify discrepancies that may be symptomatic for underlying sampling biases and related issues. This may help to highlight the problems of skewed effort distributions for habitat suitability modeling in the marine environment on very large scales. Future evaluation of RES predictions for species with available sighting data sets using standard evaluation statistics based on confusion matrices and thresholds optimized by receiver–operator curves for species presence would be helpful for a case-by-case investigation of the extent to which our predictions correspond more closely to a species' fundamental versus its realized niche.

In a management context, RES predictions can usefully supplement small-scale studies by providing some greater context of general boundaries and potential focal areas of species' occurrences in unsurveyed regions. Thus, the RES model may provide cost-efficient starting points to focus future research and survey efforts. This is especially practical when dealing with the many data-poor species in the lesser-studied regions of the world, such as some of the rare and endangered beaked whales. The usefulness of habitat prediction models to minimize anthropogenic impacts on endangered species of marine mammals through the implementation of effectively designed marine reserves has already been demonstrated on relatively small scales (Mullin et al. 1994b, Moses & Finn 1997, Hooker et al. 1999). RES modeling may be equally useful when attempting to delineate efficient marine protected areas or critical habitat on larger geographic scales, by generating global spatially explicit indexes of biodiversity and species richness, or visualizing potential geographic hotspots of high conflict with fisheries or other human operations (Kaschner 2004, K. Kaschner et al. unpubl. data).

Acknowledgements. This research was conducted as part of the 'Sea Around Us' project, initiated and funded by the Pew Charitable Trusts of Philadelphia, USA, and with support from the National Science and Engineering Research Council of Canada. A 'Li Tze Fong' graduate fellowship and a partial university graduate fellowship from the University of British Columbia provided additional support for K. Kaschner. We are grateful to K. Bodtger, E. Gregr, A. Hall, and B. Wilson (MMRU-UBC), as well as S. Jansen (BFA-Hamburg) for many fruitful discussions on methodology and model development. K. Willis helped compile habitat usage information, and M. Denoth, A. Winship, and R. Joy offered invaluable advice and assistance during the development of the validation approach. Comments from Falk Huettmann and the anonymous reviewers greatly helped to improve this manuscript. We are indebted to C. Allison and the Secretariat of the International Whaling Commission for providing us with the IWC catch and sighting data sets. We also thank P. Hammond, SMRU, for the SCANS data set and P. Olesiuk, DFO, for the northern fur seal data. In addition, we are grateful for the ice edge data provided by the United States National Snow & Ice Data Centre.

LITERATURE CITED

Angel MV (1992) Long-term, large-scale patterns in marine pelagic systems. In: Giller PS, Hildrew AG, Raffaelli DG (eds) Aquatic ecology: scale, pattern and process. Proceedings of the 34th symposium of the British Ecological Society with the American Society of Limnology and Oceanography, University College, Cork. Blackwell Scientific Publications, Boston, p 377–402

Arnold PW (2002) Irrawaddy dolphin—*Orcaella brevirostris*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 652–654

Arnould JPY, Hindell M (2001) Dive behaviour, foraging locations, and maternal attendance patterns of Australian fur seals (*Arctocephalus pusillus doriferus*). Can J Zool 79:35–48

Au DWK, Perryman WL (1985) Dolphin habitats in the eastern tropical Pacific. Fish Bull 83:623–644

Austin MP (2002) Spatial prediction of species distribution: an interface between ecological theory and statistical modeling. Ecol Model 157:101–118

Ballance LT, Pitman RL (1998) Cetaceans of the western tropical Indian Ocean: distribution, relative abundance, and comparisons with cetacean communities of two other tropical ecosystems. Mar Mamm Sci 14:429–459

Bannister JL, Brownell RLJ, Best PB, Donovan GP (2001) Report of the workshop on the comprehensive assessment of right whales: a worldwide comparison. J Cetacean Res Manage Spec Iss 2:1–60

Baumgartner MF (1997) The distribution of Risso's dolphin (*Grampus griseus*) with respect to the physiography of the northern Gulf of Mexico. Mar Mamm Sci 13:614–638

Baumgartner MF, Mullin KD, May LN, Leming TD (2001) Cetacean habitats in the northern Gulf of Mexico. Fish Bull 99:219–239

Baumgartner MF, Cole TVN, Clapham PJ, Mate B (2003) North Atlantic right whale habitat in the lower Bay of Fundy and on the SW Scotian Shelf during 1999–2001. Mar Ecol Prog Ser 264:137–154

Belcher RL, Lee TE (2002) *Arctocephalus townsendi*. Mamm Species 700:1–5

Bengtson JL, Stewart BS (1997) Diving patterns of a Ross seal (*Ommatophoca rossii*) near the east coast of the Antarctic peninsula. Polar Biol 18:214–218

Bester MN, Erickson AW, Ferguson JWH (1995) Seasonal change in the distribution and density of seals in the pack ice off Princess Martha Coast, Antarctica. Antarct Sci 7: 357–364

Bonner WN (1984) Lactation strategies in pinnipeds: problems for a marine mammalian group. Proc Symp Zool Soc Lond 51:253–272

Boyce MS, Vernier PR, Nielsen SE, Schmiegelow FKA (2002) Evaluating resource selection functions. Ecol Model 157(2–3):281–300

Boyd IL (1998) Time and energy constraints in pinniped lactation. Am Nat 152:717–728

Bradshaw CJA, Davis LS, Purvis M, Zhou Q, Benwell GL (2002) Using artificial neural networks to model the suitability of coastline for breeding by New Zealand fur seals (*Arctocephalus forsteri*). Ecol Model 148:111–131

Brierley AS, Fernandes PG, Brandon MA, Armstrong F and 8 others (2002) Antarctic krill under sea ice: elevated abundance in a narrow band just south of ice edge. Science 295(5561):1890–1892

Brownell RLJ, Best PB, Prescott JH (eds) (1983) Right whales: past and present status — Rep Int Whal Comm. Spec Issue 10, IWC, Cambridge

Buckland ST, Anderson DR, Burnham KP, Laake JL (1993) Distance sampling: estimating abundance of biological populations. Chapman & Hall, London

Campagna C, Werner R, Karesh W, Marin MR, Koontz F, Cook R, Koontz C (2001) Movements and location at sea of South American sea lions (*Otaria flavescens*). J Zool 255:205–220

Cañadas A, Sagarminaga R, García-Tiscar S (2002) Cetacean distribution related with depth and slope in the Mediterranean waters off southern Spain. Deep-Sea Res I 49: 2053–2073

Cañadas A, de Stephanis R, Sagarminaga R, Uriquiola E, Hammond PS (2003) Habitat selection models as conservation tool: proposal of marine protected areas for cetaceans in southern Spain. Proceedings of the 15th biennial conference on the biology of marine mammals. Society of Marine Mammalogy, Greensboro, p 28–29 (Abstract)

Carlström J, Denkinger J, Feddersen P, Øien N (1997) Record of a new northern range of Sowerby's beaked whale (*Mesoplodon bidens*). Polar Biol 17:459–461

Clark CW, Lamberson (1982) An economic history and analysis of pelagic whaling. Mar Policy 6(2):103–120

Compton RC (2004) Predicting key habitat and potential distribution of northern bottlenose whales (*Hyperoodon ampullatus*) in the northwest Atlantic Ocean. MRes, University of Plymouth

Corkeron PJ, Connor RC (1999) Why do baleen whales migrate? Mar Mamm Sci 15:1228–1245

Costa DP (1991) Reproductive and foraging energetics of high latitude penguins, albatrosses and pinnipeds: implications for life history patterns. Am Zool 31:111–130

Dalebout ML, Mead JG, Baker CS, Baker AN, van Helden AL (2002) A new species of beaked whale *Mesoplodon perrini* sp. n. (Cetacea: Ziphiidae) discovered through phylogenetic analysis of mitochondrial DNA sequences. Mar Mamm Sci 18:577–608

Davis RW, Fargion GS, May N, Leming TD, Baumgartner MF, Evans WE, Hansen LJ, Mullin KD (1998) Physical habitat of cetaceans along the continental slope in the north central and western Gulf of Mexico. Mar Mamm Sci 14:490–507

Davis RW, Ortega-Ortiz JG, Ribic CA, Evans WE and 6 others (2002) Cetacean habitat in the northern oceanic Gulf of Mexico. Deep-Sea Res I 49:121–142

Deecke V (2004) Update COSEWIC status report on the grey whale *Eschrichtius robustus* (Pacific population). In: COSEWIC assessment and status report on the grey whale *Eschrichtius robustus* in Canada. Committee On the Status of Endangered Wildlife In Canada, Ottawa, p iv + 36

Dellinger T, Trillmich F (1999) Fish prey of the sympatric Galapagos fur seals and sea lions: seasonal variation and niche separation. Can J Zool 77:1204–1261

Dietz R, Heide-Jørgensen MP (1995) Movements and swimming speed of narwhals, *Monodon monoceros*, equipped with satellite transmitters in Melville Bay, northwest Greenland. Can J Zool 73:2106–2119

Duguy R (1975) Contribution à l'étude des mammifères marin de la côte nord-ouest Afrique. Rev Trav Inst Pech Marit 39:321–332

Engler R, Guisan A, Rechsteiner L (2004) An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. J Appl Ecol 41:263–274

Evans PGH (1980) Cetaceans in British waters. Mamm Rev 10:1–52

Fedoseev G (2002) Ribbon seal — *Histriophoca fasciata*. In: Perrin WF, Wursig B, Thewissen HGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 1027–1030

Fielding AH, Bell JF (1997) A review of methods for the assessment of prediction errors in conservation presence/absence models. Environ Conserv 24:38–49

Findlay KP, Best PB, Ross GJB, Cockcroft VG (1992) The distribution of small odontocete cetaceans off the coasts of South Africa and Namibia. S Afr J Mar Sci 12:237–270

Folkow LP, Blix AS (1995) Distribution and diving behaviour of hooded seals. In: Blix AS, Walløe L, Ulltang Ø (eds) Whales, seals, fish and man. Proceedings of the international symposium on the biology of marine mammals in the North East Atlantic, Vol 4. Elsevier, Amsterdam, p 193–202

Folkow LP, Blix AS (1999) Diving behaviour of hooded seals (*Cystophora cristata*) in the Greenland and Norwegian Seas. Polar Biol 22:61–74

Folkow LP, Mårtensson PE, Blix AS (1996) Annual distribution of hooded seal (*Cystophora cristata*) in the Greenland and Norwegian Seas. Polar Biol 16:179–189

Forney KA, Barlow J (1998) Seasonal patterns in the abundance and distribution of California cetaceans, 1991–1992. Mar Mamm Sci 14:460–489

Freeland WJ, Bayliss P (1989) The Irrawaddy river dolphin, *Orcaella brevirostris*, in coastal waters of the Northern Territory, Australia: distribution, abundance and seasonal changes. Mammalia 53:49–58

Gales NJ, Shaughnessy PD, Dennis TE (1994) Distribution, abundance and breeding cycle of the Australian sea lion (*Neophoca cinerea*). J Zool 234:353–370

Gardner SC, Chavez-Rosales S (2000) Changes in the relative abundance and distribution of gray whales (*Eschrichtius robustus*) in Magdalena Bay, Mexico during an El Niño event. Mar Mamm Sci 16:728–738

Gaskin DE (1972) Whales, dolphins and seals, with special reference to the New Zealand region. Heinemann Educational Books, London

Gaskin DE (1991) An update on the status of the right whale *Eubalaena glacialis* in Canada. Can Field-Nat 105:198–205

Gaston KJ (1994) Measuring geographic range sizes. Ecography 17:198–205

Gilmartin WG, Forcada J (2002) Monk seals — *Monachus monachus*, *M. tropicalis* and *M. schauinslandi*. In: Perrin WF, Würsig B, Thewissen HGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 756–759

Goodall RNP (1997) Review of sightings of the hourglass dolphin, *Lagenorhynchus cruciger*, in the South American sector of the Antarctic and sub-Antarctic. Rep Int Whal Comm 47:1001–1013

Goodall RNP (2002) Hourglass dolphin—*Lagenorhynchus cruciger*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 583–585

Gregr EJ (2000) An analysis of historic (1908–1967) whaling records from British Columbia, Canada. MS, University of British Columbia, Vancouver

Gregr EJ, Trites AW (2001) Predictions of critical habitat for five whale species in the waters of coastal British Columbia. Can J Fish Aquat Sci 58:1265–1285

Griffin RB (1999) Sperm whale distributions and community ecology associated with a warmcore ring off Georges Bank. Mar Mamm Sci 15:33–52

Guinotte JM, Bartley JD, Iqbal A, Fautin DG, Buddemeier RW (2006) Modeling habitat distribution from organism occurrences and environmental data: a case study using anemonefishes and their sea anemone hosts. Mar Ecol Prog Ser (in this Theme Section)

Guisan A, Zimmermann N (2000) Predictive habitat distribution models in ecology. Ecol Model 135:147–186

Hall LS, Krausman PR, Morrison ML (1997) The habitat concept and a plea for standard terminology. Wildl Soc Bull 25:173–182

Hamazaki T (2002) Spatiotemporal prediction models of cetacean habitats in the mid-western North Atlantic Ocean (from Cape Hatteras, North Carolina, USA to Nova Scotia, Canada). Mar Mamm Sci 18:920–939

Hammond PS, Berggren P, Benke H, Borchers DL and 6 others (2002) Abundance of harbor porpoise and other cetaceans in the North Sea and adjacent waters. J Appl Ecol 39: 361–376

Heath CB (2002) California, Galapagos, and Japanese sea lions. In: Perrin WF, Würsig B, Thewissen HGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 180–186

Hedley SL, Buckland ST (2004) Spatial models for line-transect sampling. J Agric Biol Environ Stat 9:181–199

Hedley SL, Buckland ST, Borchers DL (1999) Spatial modeling from line transect data. J Cetacean Res Manage 1:255–264

Heide-Jørgensen MP (2002) Narwhal—*Monodon monoceros*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 783–787

Heide-Jørgensen MP, Dietz R, Laidre KL, Richard P, Orr J, Schmidt HC (2003) The migratory behaviour of narwhals (*Monodon monoceros*). Can J Zool 81:1298–1305

Hewitt RP, Lipsky JD (2002) Krill. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 676–684

Hirzel AH, Guisan A (2002) Which is the optimal sampling strategy for habitat suitability modelling? Ecol Model 157(2–3):331–341

Hooker SK, Whitehead H, Gowans S (1999) Marine protected area design and the spatial and temporal distribution of cetaceans in a submarine canyon. Conserv Biol 13:592–602

Huettmann F, Diamond AW (2001) Seabird colony locations and environmental determination of seabird distribution: a spatially explicit breeding seabird model for the Northwest Atlantic. Ecol Model 141:261–298

Hutchinson GE (1957) Concluding remarks. Cold Spring Harbor Quant Biol 22:415–427

IWC (International Whaling Commission) (2001a) IDCR-DESS SOWER survey data set (1978–2001). IWC, Cambridge

IWC (International Whaling Commission) (2001b) IWC catch database (1800–1999). IWC, Cambridge

Jaquet N (1996) How spatial and temporal scales influence understanding of sperm whale distribution: a review. Mamm Rev 26:51–65

Jaquet N, Whitehead H (1996) Scale-dependent correlation of sperm whale distribution with environmental features and productivity in the South Pacific. Mar Ecol Prog Ser 135:1–9

Jefferson TA, Karczmarski L (2001) *Sousa chinensis*. Mamm Species 655:1–9

Jefferson TA, Newcomer MW (1993) *Lissodelphi borealis*. Mamm Species 425:1–6

Jefferson TA, Schiro AJ (1997) Distribution of cetaceans in the offshore Gulf of Mexico. Mamm Rev 27:27–50

Jefferson TA, Leatherwood S, Webber MA (1993) Marine mammals of the world. FAO, Rome

Jefferson TA, Newcomer MW, Leatherwood S, van Waerebeek K (1994) Right whale dolphins *Lissodelphis borealis* (Peale, 1848) and *Lissodelphis peronii* (Lacepede, 1804). In: Ridgway SH, Harrison RJ (eds) The first book of dolphins—handbook of marine mammals, Vol 5. Academic Press, San Diego, CA, p 335–362

JMP (?) (2000) Statistics and graphics guide, Ver 4. SAS Institute Cary, NC

Johnson DH (1980) The comparison of usage and availability measurements for evaluating resource preference. Ecology 61:65–71

Jones ML, Swartz SL (2002) Gray whale—*Eschrichtius robustus*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 524–536

Karczmarski L, Cockcroft VG, McLachlan A (2000) Habitat use and preferences of Indo-Pacific humpback dolphins *Sousa chinensis* in Algoa Bay, South Africa. Mar Mamm Sci 16: 65–79

Kasamatsu F, Joyce GG (1995) Current status of odontocetes in the Antarctic. Antarc Sci 7:365–379

Kasamatsu F, Hembree D, Joyce G, Tsunoda LM, Rowlett R, Nakano T (1988) Distributions of cetacean sightings in the Antarctic: results obtained from the IWC/IDCR minke whale assessment cruises 1978/79–1983/84. Rep Int Whal Comm 38:449–487

Kasamatsu F, Matsuoka K, Hakamada T (2000) Interspecific relationships in density among the whale community in the Antarctic. Polar Biol 23:466–473

Kaschner K (2004) Modeling and mapping of resource overlap between marine mammals and fisheries on a global scale. PhD thesis, University of British Columbia, Vancouver

Kasuya T, Miyashita T (1997) Distribution of Baird's beaked whales off Japan. Rep Int Whal Comm 47:963–968

Kenney RD (2002) North Atlantic, North Pacific, and southern right whales—*Eubalaena glacialis*, *E. japonica*, and *E. australis*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 806–813

Kenney RD, Winn HE (1986) Cetacean high-use habitats of the northeast United States continental shelf. Fish Bull 84:345–357

Kenyon KW (1981) Monk seals—*Monachus* (Fleming, 1822). In: Ridgway SH, Harrison RJ (eds) Seals—handbook of marine mammals, Vol 2. Academic Press, London, p 195–220

Knowlton AR, Sigurjónsson J, Ciano JN, Kraus SD (1992) Long-distance movements of North Atlantic right whales (*Eubalaena glacialis*). Mar Mamm Sci 8:397–405

Knox GA (1994a) The biology of the southern ocean. Cambridge University Press, Cambridge

Knox GA (1994b) Seals. In: Knox GA (ed) The biology of the southern ocean. Cambridge University Press, Cambridge, p 141–160

Kovacs KM, Lavigne DM (1986) *Cystophora cristata*. Mamm Species 258:1–9

Lalas C, Bradshaw CJA (2001) Folklore and chimerical numbers: review of a millennium of interaction between fur seals and humans in the New Zealand region. NZ J Mar Freshw Res 35:477–497

Lander ME, Gulland FMD, DeLong RL (2000) Satellite tracking a rehabilitated Guadalupe fur seal (*Arctocephalus townsendi*). Aquat Mamm 26:137–142

LeDuc RG (2002) Biogeography. In: Perrin WF, Würsig B, Thewissen HGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 99–102

Levins R (1966) The strategy of model building in population biology. Am Sci 54:421–431

Ling JK (1992) *Neophoca cinerea*. Mamm Species 392:1–7

Ling JK (2002) Australian sea lion—*Neophoca cinerea*. In: Perrin WF, Wursig B, Thewissen HGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 51–54

MacLeod CD (2005) Niche partitioning, distribution and competition in North Atlantic beaked whales. PhD, University of Aberdeen, UK

Manel S, Williams HC, Ormerod SJ (2001) Evaluating presence/absence models in ecology: the need to account for prevalence. J Appl Ecol 38:921–931

Manly BFJ, McDonald LL, Thomas DL, McDonald TL, Erickson WP (2002) Resource selection by animals: statistical design and analysis for field studies. Kluwer Academic, Dordrecht

Martin AR, Reeves RR (2002) Diversity and zoogeography. In: Hoelzel AR (ed) Marine mammal biology—an evolutionary approach. Blackwell Science, London, p 1–37

Martin AR, Kingsley MCS, Ramsay MA (1994) Diving behaviour of narwhals (*Monodon monoceros*) on their summer grounds. Can J Zool 72:118–125

Matsuoka K, Ensor P, Hakamada T, Shimada H, Nishiwaki S, Kasamatsu F, Kato H (2003) Overview of minke whale sightings surveys conducted on IWC/IDCR and SOWER Antarctic cruises from 1978/79 to 2000/01. J Cetacean Res Manage 5:173–201

Mitchell E, Mead JG (1977) The history of the gray whale in the Atlantic Ocean. In: Proceedings of the 2nd conference on the biology of marine mammals. Society of Marine Mammalogy, San Diego, CA, p 12

Mitchell E, Kozicki VM, Reeves RR (1983) Sightings of right whales *Eubalaena glacialis* on the Scotian shelf, 1966–1972. In: Brownell RLJ, Best PB, Prescott JH (eds) Right whales: past and present status—Rep Int Whal Comm. Spec Issue 10, IWC, Cambridge, p 83–108

Mizuno AW, Wada A, Ishinazaka T, Hattori K, Watanabe Y, Ohtaishi N (2002) Distribution and abundance of spotted seals (*Phoca largha*) and ribbon seals (*Phoca fasciata*) in the southern Sea of Okhotsk. Ecol Res 17:79–96

Moisen GG, Frescino TS (2002) Comparing five modeling techniques for predicting forest characteristics. Ecol Model 157:209–225

Moore SE (2000) Variability of cetacean distribution and habitat selection in the Alaskan Arctic, autumn 1982–1991. Arctic 53:448–460

Moore SE, DeMaster DP (1997) Cetacean habitats in the Alaskan Arctic. J Northwest Atl Fish Sci 22:55–69

Moore SE, Waite JM, Friday NA, Honkalehto T (2002) Cetacean distribution and relative abundance on the central-eastern and the southeastern Bering Sea shelf with reference to oceanographic domains. Prog Oceanogr 55: 249–261

Mörzer Bruyns WFJ (1971) Field guide of whales and dolphins. Tor, Amsterdam

Moses E, Finn JT (1997) Using geographic information systems to predict North Atlantic right whale (*Eubalaena glacialis*) habitat. J Northwest Atl Fish Sci 22:37–46

Mullin KD, Higgins LV, Jefferson TA, Hansen LJ (1994a) Sightings of the Clymene dolphin (*Stenella clymene*) in the Gulf of Mexico. Mar Mamm Sci 10:464–470

Mullin KD, Hoggard W, Roden CL, Lohoefener RR, Rogers CM, Taggart B (1994b) Cetaceans on the upper continental slope in the north-central Gulf of Mexico. Fish Bull 92: 773–786

Murase H, Matsuoka K, Ichii T, Nishiwaki S (2002) Relationship between the distribution of euphausiids and baleen whales in the Antarctic (35° E–145° W). Polar Biol 25: 135–145

NOAA/NODC (National Oceanic and Atmospheric Administration/National Oceanographic Data Center) (1998) World ocean atlas 1998. Ocean Climate Laboratory, NODC, Washington, DC

Oksanen J, Minchin PR (2002) Continuum theory revisited: What shape are species responses along ecological gradients? Ecol Model 157:119–129

Parra GJ, Azuma C, Preen AR, Corkeron PJ, Marsh H (2002) Distribution of Irrawaddy dolphins, *Orcaella brevirostris*, in Australian waters. Raffles Bull Zool Suppl 10:141–154

Parrish FA, Craig MP, Ragen TJ, Marshall GJ, Buhleier BM (2000) Identifying diurnal foraging habitat of endangered Hawaiian monk seals using a seal-mounted video camera. Mar Mamm Sci 16:392–412

Parrish FA, Abernathy K, Marshall GJ, Buhleier BM (2002) Hawaiian monk seals (*Monachus schauinslandi*) foraging in deep-water coral beds. Mar Mamm Sci 18:244–258

Payne PM, Heinemann DW (1993) The distribution of pilot whales (*Globicephala* spp.) in shelf/shelf-edge and slope waters of the northeastern United States, 1978–1988. In: Donovan GP, Lockyer CH, Martin AR (eds) Biology of northern hemisphere pilot whales—Rep Int Whal Comm. Spec Issue 14, IWC, Cambridge, p 51–68

Perrin WF, Brownell RLJ (2002) Minke whales—*Balaenoptera acutorostrata* and *B. bonaerensis*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 750–754

Perrin WF, Mitchell ED, Mead JG, Caldwell DK, van Bree PJH (1981) *Stenella clymene* a rediscovered tropical dolphin of the Atlantic. J Mamm 62:583–598

Perrin WF, Leatherwood S, Collet A (1994) Fraser's dolphin *Lagenodelphis hosei* (Fraser, 1956). In: Ridgway SH, Harrison RJ (eds) The first book of dolphins—handbook of marine mammals, Vol 5. Academic Press, San Diego, CA, p 225–240

Perrin WF, Würsig B, Thewissen JGM (eds) (2002) Encyclopedia of marine mammals. Academic Press, San Diego, CA

Perry SL, DeMaster DP, Silber GK (1999) The status of endangered whales. Mar Fish Rev 61:1–74

Peterson AT (2001) Predicting species' geographic distributions based on ecological niche modeling. Condor 103:599–605

Peterson AT, Navarro-Sigüenza AG (1999) Alternate species concepts as bases for determining priority conservation areas. Conserv Biol 13:427–431

Peterson AT, Egbert SL, Sánchez-Cordero V, Price KP (2000) Geographic analysis of conservation priority: endemic birds and mammals in Veracruz, Mexico. Biol Conserv 93:85–94

Pitman RL, Ensor P (2003) Three forms of killer whales (*Orcinus orca*) in Antarctic waters. J Cetacean Res Manage 5:131–139

Pitman RL, Ballance LT, Mesnick SI, Chivers SJ (2001) Killer whale predation on sperm whales: observations and implications. Mar Mamm Sci 17:494–507

Platt T, Sathyendranath S (1992) Scale, pattern and process in marine ecosystems. In: Giller PS, Hildrew AG, Raffaelli DG (eds) Aquatic ecology: scale, pattern and process. Proceedings of the 34th symposium of the British Ecological Society with the American Society of Limnology and Oceanography, University College, Cork. Blackwell Scientific Publications, Boston, MA, p 593–599

Ponganis PJ, Gentry RL, Ponganis EP, Ponganis KV (1992) Analysis of swim velocities during deep and shallow dives of two northern fur seals *Callorhinus ursinus*. Mar Mamm Sci 8:69–75

Reijnders P, Brasseur S, van der Toorn J, van der Wolf R, Boyd IL, Harwood J, Lavigne DM, Lowry L (1993) Seals, fur seals, sea lions, and walrus. Status survey and conservation action plan. IUCN/SSC Specialist Group, International Union for the Conservation of Nature and Natural Resources, Gland

Ribic CA, Ainley DG, Fraser WR (1991) Habitat selection by marine mammals in the marginal ice zone. Antarct Sci 3:181–186

Rice DW (1998) Marine mammals of the world—Systematics and distribution. Spec Publ 4, Allen Press, Lawrence, KS

Ridgway SH, Harrison RJ (eds) (1981a) The walrus, sea lions, fur seals and sea otter—handbook of marine mammals, Vol 1. Academic Press, London

Ridgway SH, Harrison RJ (eds) (1981b) Seals—handbook of marine mammals, Vol 2. Academic Press, London

Ridgway SH, Harrison RJ (eds) (1985) The sirenians and baleen whales—handbook of marine mammals, Vol 3. Academic Press, London

Ridgway SH, Harrison RJ (eds) (1989) The river dolphins and the larger toothed whales—handbook of marine mammals, Vol 4. Academic Press, San Diego, CA

Ridgway SH, Harrison RJ (eds) (1994) The first book of dolphins—Handbook of marine mammals, Vol 5. Academic Press, San Diego, CA

Ridgway SH, Harrison RJ (eds) (1999) The second book of dolphins and the porpoises—Handbook of marine mammals, Vol 6. Academic Press, New York

Rosenbaum HC, Brownell RL Jr, Brown MW, Schaeff C and 16 others (2000) World-wide genetic differentiation of *Eubalaena*: questioning the number of right whale species. Mol Ecol 9:1793–1802

Ross GJB (2002) Humpback dolphins—*Sousa chinensis*, *S. plumbea*, and *S. teuszi*. In: Perrin WF, Würsig B, Thewissen JGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 585–589

Rugh DJ, Muto MM, Moore SE, DeMaster DP (1999) Status review of the eastern North Pacific stock of gray whales. NOAA Tech Mem NMFS-AFSC 103:1–93

Schmelzer I (2000) Seals and seascapes: covariation in Hawaiian monk seal subpopulations and the oceanic landscape of the Hawaiian Archipelago. J Biogeogr 27:901–914

Shao G, Halpin PN (1995) Climatic controls of eastern North American coastal tree and shrub distributions. J Biogeogr 22:1083–1089

Sigurjónsson (1995) On the life history and autecology of North Atlantic rorquals. In: Whales, seals, fish and man—Proc Int Symp Biology of Matine Mammals in the Northeast Atlantic. Tromsø, Norway, 29 November – 1 December 1994. Elsevier, Amsterdam

Skov F, Svenning JC (2004) Potential impact of climatic change on the distribution of forest herbs in Europe. Ecography 27:366–380

Smith RC, Dustan P, Au D, Baker KS, Dunlap EA (1986) Distribution of cetaceans and sea-surface chlorophyll concentrations in the California currrent. Mar Biol 91:385–402

Splettstoesser JF, Gavrilo M, Field C, Field C, Harrison P, Messick M, Oxford P, Todd FS (2000) Notes on Antarctic wildlife: Ross seals *Ommatophoca rossii* and emperor penguins *Aptenodytes forsteri.* NZ J Zool 27:137–142

Stacey PJ (1996) Natural history and conservation of Irrawaddy dolphins, *Orcaella brevirostris*, with special reference to the Mekong River of Laos PDR. University of Victoria

Stockwell DRB, Noble IR (1992) Induction of sets of rules from animal distribution data: a robust and informative method of data analysis. Math Comput Simulation 33:385–390

Stockwell DRB, Peterson AT (2001) Controlling bias during predictive modeling with museum data. In: Scott JM, Heglund PJ, Morrison M, Raphael M, Haufler J, Wall B, Samson F (eds) Predicting plant, animal and fungi occurrences: issues of scale and accuracy. Island Press, Clovelo

Stockwell DRB, Peterson AT (2002) Effects of sample size on accuracy of species distribution models. Ecol Model 148:1–13

Store R, Jokimäki J (2003) A GIS-based multi-scale approach to habitat suitability modeling. Ecol Model 169:1–15

Thomas JA (2002) Ross seal—*Ommatophoca rossii*. In: Perrin WF, Würsig B, Thewissen HGM (eds) Encyclopedia of marine mammals. Academic Press, San Diego, CA, p 1053–1055

Thomas RM, Schulein FH (1988) The shoaling behavior of pelagic fish and the distribution of seals and gannets off Namibia as deduced from routine fishing reports 1982–1985. S Afr J Mar Sci 7:179–192

Thompson D, Duck CD, McConnell BJ, Garrett J (1998) Foraging behaviour and diet of lactating female southern sea lions (*Otaria flavescens*) in the Falkland Islands. J Zool 246:135–146

Turner MG, Gardner RH, O'Neill RV (1995) Ecological dynamics at broad scale. BioScience Suppl (Sci Biodiversity Policy):29–35

Tynan CT, DeMaster DP, Peterson WT (2001) Endangered right whales on the southeastern Bering Sea shelf. Science 294:1894

Walters C (2003) Folly and fantasy in the analysis of spatial catch rate data. Can J Fish Aquat Sci 60:1433–1436

Waring GT, Quintal JM, Fairfield CP, Clapham PJ and 10 others (2002) US Atlantic and Gulf of Mexico marine mammal stock assessments—2002. NOAA Tech mem, Report No. NMFS-NE-169, US Department of Commerce, Washington, DC

Watson R, Kitchingman A, Gelchu A, Pauly D (2004) Mapping global fisheries: sharpening our focus. Fish Fisheries Ser 5: 168–177

Weller DW, Reeve SR, Burdin AM, Wuersig B, Brownell RLJ (2002) A note on the spatial distribution of western gray whales (*Eschrichtius robustus*) off Sakhalin Island, Russia in 1998. J Cetacean Res Manage 4:13–17

Werner R, Campagna C (1995) Diving behaviour of lactating southern sea lions (*Otaria flavescens*) in Patagonia. Can J Zool 73:1975–1982

Whitehead H, Jaquet N (1996) Are the charts of Maury and Townsend good indicators of sperm whale distribution and seasonality? Rep Int Whal Comm 46:643–647

Woodley TH, Gaskin DE (1996) Environmental characteristics of North Atlantic right and fin whale habitat in the lower Bay of Fundy, Canada. Can J Zool 74:75–84

Yen PPW, Sydeman WJ, Hyrenbach KD (2004) Marine bird and cetacean associations with bathymetric habitats and shallow-water topographies: implications for trophic transfer and conservation. J Mar Syst 50:79–99

Zaniewski AE, Lehmann A, Overton JM (2002) Predicting, species spatial distributions using presence-only data: a case study of native New Zealand ferns. Ecol Model 157:261–280

Zar JH (1996) Biostatistical analysis. Prentice Hall, Upper Saddle River, NJ

Editorial responsibility: Howard I. Browman (Associate Editor-in-Chief), Storebø, Norway

Submitted: March 18, 2005; Accepted: September 21, 2005
Proofs received from author(s): May 17, 2006