



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

Mark J. Costello^{1,*}, Edward Vanden Berghe²

¹Leigh Marine Laboratory, University of Auckland, PO Box 349, Warkworth, New Zealand

²Flemish 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

—Resale or republication not permitted without written consent of the publisher—

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 com-

plement 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 (Leonart 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.gisnetwork.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 $\leq 0.5\%$ 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

<p>Data publication</p> <ul style="list-style-type: none"> • 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 <p>Consequences</p> <ul style="list-style-type: none"> • 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 <ul style="list-style-type: none"> • 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 <ul style="list-style-type: none"> • 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 <ul style="list-style-type: none"> • 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

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 proto-

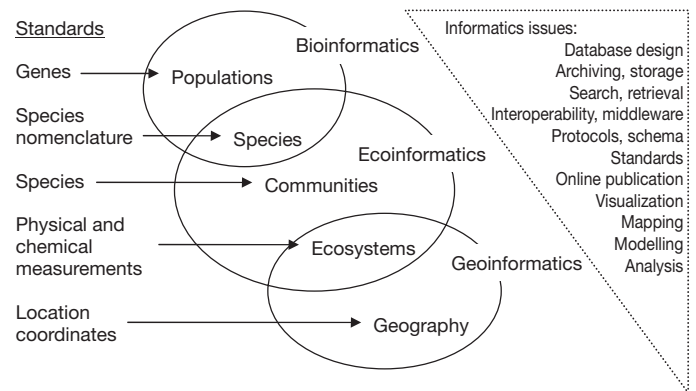


Fig. 1. Diagram showing how 3 sub-components of biodiversity informatics relate to aspects of biodiversity from genes to ecosystems and the environment. Each component has 1 unique aspect and also areas that overlap with others. The essential and standard fields for each aspect and informatics issues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

cols, 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 & Wicczorek 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 in-

cludes 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 km² (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, semi-automated '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 invest-

ments 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 non-governmental 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

<p>Science culture</p> <ol style="list-style-type: none"> 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 <p>Informatics</p> <ol style="list-style-type: none"> 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 <p>Data available</p> <ol style="list-style-type: none"> 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 <p>Consequences for efficiency in science</p> <ol style="list-style-type: none"> 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, <i>in situ</i> 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 publica-

tion 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, Cachueta-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 News* 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 multi-disciplinary 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 multi-dimensional 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 CSRP 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&fcid=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, Agosin 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 Instituut 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*