INTEGRATING BIOLOGICAL DATA INTO OCEAN OBSERVING SYSTEMS: THE FUTURE ROLE OF OBIS

Edward Vanden Berghe[1], Patrick N. Halpin[2], Fabio Lang da Silveira[3], Karen Stocks[4], Fred Grassle[5]

[1] *International OBIS (Ocean Biogeographic Information System) Secretariat, Rutgers University. 71 Dudley Road, New Brunswick, NJ 08901, USA, Email:* evberghe@iobis.org
[2] *Marine Geospatial Ecology Lab, Duke University, 450 Research Dr., Durham, NC 27708-0001, USA, Email:* phalpin@duke.edu
[3] *Instituto de Biociencias, University of Sao Paulo. Rua do Matao trav. 14 #111, Citade Universitaria 05508-900 Sao Paulo, Brazil, Email:* fldsilve@usp.br
[4] *San Diego Supercomputer Center, University of California, San Diego. 9500 Gilman Drive MC0505, La Jolla, CA 92093, USA, Email:* kstocks@sdsc.edu
[5] *Institute of Marine and Coastal Sciences, Rutgers University. 71 Dudley Road, New Brunswick, NJ 08901, US, Email:* grassle@marine.rutgers.edu

ABSTRACT

The future data needs of ocean science and ocean resource management will require a more seamless and accessible coupling of biological data with physical oceanographic processes. This bio-physical data framework will be built through the active integration of data from an extensive variety of sensors, observers, platforms and data archives across a wide range of space and time scales. This necessary synthesis of raw biological data into useful information and potentially new understanding is dependent on both new developments in ocean exploration as well as developments in information systems and informatics. The Ocean Biogeographic Information System (OBIS) is poised to play a significant and expanding role in the evolving ocean observation system [1] and [2].

OBIS was created as the data integration component of the Census of Marine Life [3], [4], and [5]. From the start OBIS was conceived to be a global and distributed system, giving control of data to data providers [6], with strong ties to existing national and international biodiversity information systems [2] and [6]. Since then, OBIS has developed as a facility for finding, absorbing, integrating, assessing and analysing data about life in the oceans. The maps in Fig. 1 and Fig. 2 illustrate the global nature of OBIS. OBIS is aimed at stimulating research and generating new hypotheses on evolutionary processes and species distributions. It serves as a basis for informed management of marine biodiversity by making data freely accessible. It integrates data from many sources, over a wide range of marine themes, from poles to the equator, from microbes to whales. It is the largest provider of marine species distribution information, and one of the largest contributors to the Global Biodiversity Information Facility (GBIF).

Institutionally, OBIS is growing rapidly as a distributed system with Regional OBIS Nodes (RONs) in all continents. Each RON is self-sustained and is the geographical backbone for further development of OBIS data content. The institutes hosting the RONs are an asset for OBIS as a network and have proven to be very supportive of OBIS activities and objectives. Keeping this network functional is a priority for the future of OBIS.
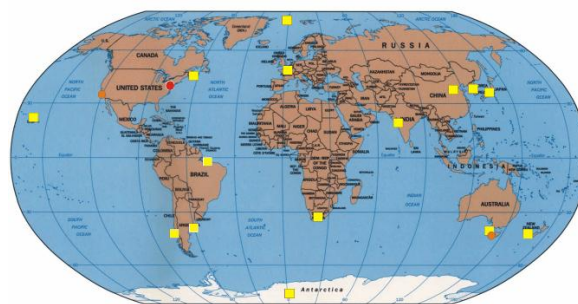

Figure 1. Locations of the offices of Regional OBIS Nodes

In addition to the Regional Nodes, OBIS has thematic nodes for major subsets of marine life. OBIS SEAMAP, the repository for data on marine birds, turtles, and mammals, is developing new ways to visualize migrations of these animals and to understand their habitats [7 and 8]. The Biogeoinformatics of Hexacorals website maintains an authoritative, global anemone and coral database [9]. FishBase, who partnered with OBIS from its inception, contains comprehensive information on fishes [10]. The OBIS micro-organisms component (MICROBIS) is breaking completely new ground by defining the known world of micro-organisms using new molecular bar-coding approaches to define microbial taxa. The Continuous Plankton Recorder (CPR), managed by the Sir Alex Hardy Foundation of Oceanographic Science (SAHFOS) provides a unique and very large dataset. One of the strengths of the CPR data is that it has been collected in a standard way for more than half a century [11].

OBIS strives to be interoperable with other initiatives. It uses the Distributed Generic Information Retrieval (DiGIR, http://digir.net; [12]) protocol as its data

exchange standard, and the OBIS Schema, a derivative of the Darwin Core, as its format specification for content [13]. Through this combination, data are made available to several organisations, including GBIF. However, DiGIR is now seen by many as too restrictive. The biodiversity community was an early adopted of XML technology in exchange data and building distributed systems. Technology has evolved, and DiGIR was not adopted outside the biodiversity community. There is an urgent need to build alternatives for data exchange, based on Open Geospatial Consortium standards. This will enhance integration of biodiversity data with other ocean data, such as physical and chemical oceanography.
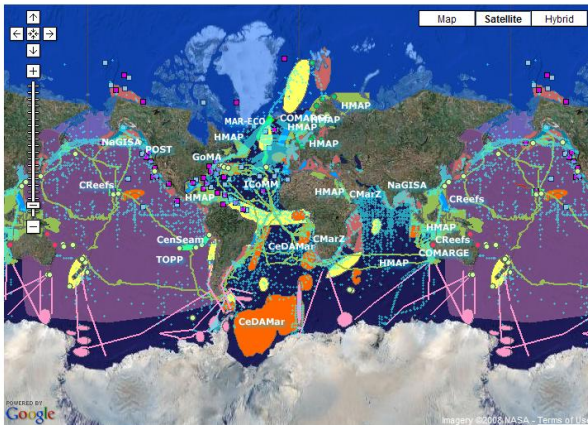

Figure 2. Map illustrating the geographic scope of the Census of Marine Life field projects

Standards development and implementation are an integral part of the activities. Data integration on a massive scale, as practiced by OBIS, is critically dependent on appropriate standards for data exchange and documentation. When possible, existing standards are used in building the OBIS infrastructure. Where needed, OBIS plays an active role in furthering standard development. The World Register of Marine Species (WoRMS, http://marinespecies.org) was developed by the OBIS community, led by its European node, as a continuously updated expert-controlled vocabulary for species names, an essential component of biodiversity data integration and quality assurance. Other quality control tools, e.g. for detecting outliers in environmental space, are being developed.

The international OBIS portal (http://www.iobis.org) is developed and maintained at Rutgers University, with assistance from many in the community [13] and [14]. It offers an on-line, user-friendly search interface to all OBIS data. It offers facilities for downloading data, and provides access to several external tools for data visualisation and analysis. The web site is highly used, with over a million hits per month, and more than 100,000 records viewed or downloaded per day. In April 2009, the integrated data contained 18.5 million records, from 633 distinct datasets, and 105,000 species.

With this, OBIS is by far the largest provider of primary data on species distributions, with the largest number of marine species. Researchers are using it to test different analyses of marine life distribution (e.g. [15] and [16]).

Many more datasets exist than the ones that are available through OBIS. One of the tasks we have set ourselves is to expand the holdings, and work with data custodians to make their data available for re-use. Data are extremely valuable, and in view of the historic nature of environmental observations, unique. OBIS works with marine scientists worldwide to mobilise data, including targeted campaigns of data archaeology and rescue; the cost of recovering data is only a fraction of the cost of data collection [17]. In doing so, OBIS tries to minimize the effort needed on behalf of the data provider. Often data providers are researchers collecting the data, and time spent on publishing data through OBIS should not detract from time available to generate data. It is hoped that, by publishing data through OBIS and the increased visibility this brings, data providers will have increased potential to generate funding for future work, and for collaborative ventures with others who detected their data through OBIS. Also, many data providers are under contractual or even legal obligation to make their data publicly available. It is clear that OBIS can assist in this task, and so, rather than taking away time for data collection from the data custodians, would free up time.

One of the poorly-represented information elements in OBIS is quantitative information on species abundance. The fields for this information exist in the exchange standard, but are poorly populated. Also, the way the information is defined in the OBIS Schema does not allow integration across datasets to the same extent as what is possible for presence-only information. Solving, or at least alleviating, this problem will be one of the main points of future attention for OBIS. By making quantitative information as readily available as the presence data, the information content will be of much more value to the ocean observing community.

Biological data management has its very specific challenges, over and above the challenges that data managers in other domains are confronted with. First of all, there is a lack of standardization of the measured quantities in biology and ecology. This lack of standardization stems partly from the fact that there is no single 'good' way to make a measurement. How, for example, would one measure wet weight of a recently sampled large jellyfish? The measurement will depend as much on the time the specimen spent out of the water, as on any feature with biological reality. Obviously, this should not be seen as an excuse not to make biological measurements or to share this type of data, but an argument to include all relevant metadata, and the precise protocol used to make an observation.

Taxonomy brings in an extra dimension to biological data, as compared to physical-chemical oceanography. Characteristics of individual species have a profound effect on the way the carbon or nutrients locked up in it behave in the environment. In order to understand the biology, and to be able to build predictive models of say algal blooms, we have to know which species we are dealing with [18]. Similarly, bioactive substances can be present in one species but not in another, closely related one [19]. And if we want to quantify biodiversity, it is important to discriminate between species.

Despite the importance of taxonomic names and the role they play in understanding marine biology, no complete list of species names exists. The Catalogue of Life (http://www.catalogueoflife.org/) is the main effort within the biological community to create an global authoritative list of all species, terrestrial as well as marine. The World Register of Marine Species (WoRMS, http://www.marinespecies.org) aims to become the comprehensive and guide to names of marine organisms, and feeds into the Catalogue of Life. Lack of a standard reference for taxonomic names poses a serious problem, in view of the dynamic nature of taxonomic names, and in view of the many synonyms and homonyms that exist. For example, the name *Turbinaria* can refer to a genus of green algae or to a genus of corals. The sperm whale, *Physeter macrocephalus*, has also been referred to under many different names, 22 of which are listed on the WoRMS web site. Again, these idiosyncrasies of biological data create complications for data management. It is clear that there should be documentation of the source of the taxonomic name, so that we can disambiguate homonyms. Since taxonomic names are dynamic, and can change to reflect changing scientific understanding about relationships between organisms, we have to make provisions in our biological data management systems to keep track of historical names.

Identification, assigning a taxonomic name to a biological specimen, is for many biological groups a specialized undertaking, requiring years of training. In many cases, it requires certification – for example if identifications and the resulting names are used in the framework of Environmental Impact assessment. Many biologists feel a greater sense of ownership because of the larger intellectual effort required to make a taxonomically-resolved biological observation. We'll leave it to our colleagues in physical and chemical oceanography to decide whether this is justified or not, but it certainly has the effect that the practice of sharing data within the biological community is lagging behind, compared to other oceanographic disciplines. While the previous complications mentioned require extra technical work, the reluctance to share data is a sociological one, and will need more time to be completely resolved [20]. We hope that initiatives such as GBIF and OBIS can clearly demonstrate the scientific gains of sharing data, and convince more biologists to make their data publicly available.

Below are two maps illustrating possible applications of the OBIS data. Figure 3 represents an analysis of the spread of an invasive species, *Pterois volitans* (Linnaeus, 1758), and can be used by environmental managers to plan for action. Figure 4 is a global analysis of marine biodiversity patterns (updated from [21]). Such a global map is only made possible by the massive data integration of OBIS.
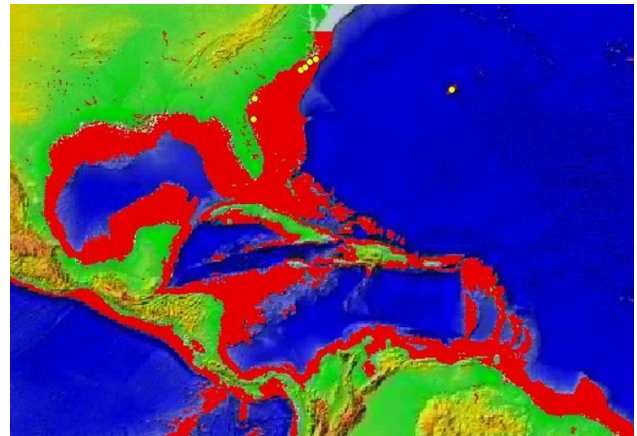

Figure 3. Observations and potential spread of an invasive species, the lionfish. Yellow dots are observations; the red area indicates where oceanographic conditions are similar to its native Red Sea, and so potential spread.

Currently, OBIS is funded by the Alfred P. Sloan Foundation. From early 2010, only partial support will be available from the Sloan Foundation, and after 2010 permanent sustained funding will need to be found from other sources. OBIS will have to rely on project funding to continue its activities. There is, however, a fundamental mismatch between OBIS' ambition to be a permanent component of the international scientific infrastructure, and the very temporary and unsure nature of project funding. For this reason, OBIS started discussions with the Intergovernmental Oceanographic Commission (IOC) of UNESCO. In its General Assembly of June 2009 the IOC accepted OBIS as one of its activities, under the International Oceanographic Data and Information Exchange (IODE) programme. Further discussions are now held to implement the integration of OBIS within IODE. One of the aspects that have to be resolved is the relationship between OBIS and the Global Ocean Observing System – how OBIS should develop in order to become the biodiversity component of GOOS. Discussions with the GOOS team at IOC, but also with the regional GOOS activities in the USA and elsewhere, will assist in defining priorities.
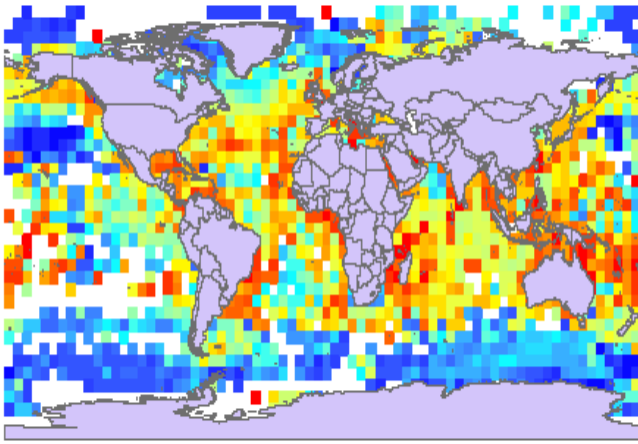
Figure 4. Hurlbert's index, ES(50), calculated on a grid of 5x5 degrees. ES(50) is the expected number of distinct species in a random sample of 50 observations; it is a bias-independent estimator of species richness. Red is high diversity, blue low, white for squares with less than 50 observations

REFERENCES

1. Poloczanska, E., Hobday, A.J., Richardson, A.J. (2008). Global database is needed to support adaptation science. *Nature* **453:** 720.

2. Grassle, J.F. (2005). Appendix 7. Biological data considerations. Pp 297-304 'Data management and communications plan for research and operational integrated ocean observing systems. Ocean.us publication 6. Washington DC.

3. Grassle, J.F., Stocks, K.I. (1999). A Global Ocean Biogeographic Information System (OBIS) for the Census of Marine Life. *Oceanography* **12(3):** 12-14.

4. Grassle, J.F. (2000). The Ocean Biogeographic Information System (OBIS): an on-line, worldwide atlas for accessing, modelling and mapping marine biological data in a multidimensional geographic context. *Oceanography* **13(3)**: 5-7.

5. Yarincik, K., O'Dor, R. (2005). The Census of Marine Life: goals, scope and strategy. *Sci. Mar.* **69 (suppl. 1):** 201-208.

6. Fornwall, M. (2000). Planning for OBIS: examining relationships with existing national and international biodiversity information systems. *Oceanography* **13(3):** 31-38.

7. Halpin P.N., Read A.J., Best B.D., Hyrenbach K.D., Fujioka E., Coyne M.S., Crowder L.B., Freeman S.A., Spoerri C. (2006). OBIS-SEAMAP: developing a biogeographic research data commons for the ecological studies of marine mammals, seabirds, and sea turtles. *Marine Ecological Progress Series* **316:** 239-246

8. Halpin, P.N., Read, A.J., Best, B.D., Fujioka, E., Donnelly, B., Kot, C., Hazen, L.J., Hyrenbach, D. &

9. Crowder, L. (2009) OBIS-SEAMAP 2.0: Developing a research data commons for the ecological studies of marine mammals, seabirds and seaturtles. *Oceanography* **22(2):** 104-115.

10. Fautin, D. (2000). Electronic Atlas of Sea anemones: an OBIS pilot project. *Oceanography* **13:** 66-69.

11. Froese, R. & Pauly, D. (eds)(2009). FishBase. World Wide Web electronic publication: www.fishbase.org, version (09/2009).

12. Reid, P. & Co-Authors (2010). "A Global Continuous Plankton Recorder Programme" in these proceedings (Vol. 2), doi:10.5270/OceanObs09.cwp.73.

13. Blum, S., Vieglais D. & Schwartz, P.J. (2001). DiGIR – Distributed Generic Information Retrieval. PowerPoint presentation, available at http://digir.sourceforge.net/events/20011106/DiGIR.ppt .

14. Rees, T., Zhang, Y. (2007). Evolving concepts in the architecture and functionality of OBIS, the Ocean Biogeographic Information System, *in*: Vanden Berghe, E. *et al.* (ed.) (2007). *Proceedings of 'Ocean Biodiversity Informatics': an international conference on marine biodiversity data management Hamburg, Germany, 29 November - 1 December, 2004. IOC Workshop Report,* 202, VLIZ Special Publication 37: pp. 167-176.

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

16. Mora, C., Tittensor, D.P. & Myers, R.A. (2008). The completeness of taxonomic inventories for describing the global diversity and distribution of marine fishes. *Proc R Soc B* **275**(**1631):** 149-155.

17. Cavalcanti, M.J. & Gallo V. (2008). Panbiogeographical analysis of distribution patterns in hagfishes (Craniata: Myxinidae). *Journal of Biogeography* **35(7):** 1258-1268.

18. Zeller, D., Froese, R. & Pauly, D. (2005). On losing and recovering fisheries and marine science data. *Marine policy* **29:** 69-73.

19. Amato, A., Kooistra, W.H.C.F., Ghiron, J.H.L., Mann, D.G., Proschold, T. & Montesor, M (2007). Reproductive isolation among sympatric cryptic species in marine diatoms. *Protist* **158:** 193-207.

20. Bryn, N. (2009). Empty archives. *Nature* **461:** 160-63.

21. Fautin, D.G. (1988). *Biomedical importance of marine organisms*. Memoirs of the California Academy of Sciences 13.

22. Grassle, J.F. & Vanden Berghe, E. (2009). Census of Marine Life 2010. P 63. In: Earle, S.A. & L.K. Glover. '*Ocean – an illustrated atlas*'. National Geographic, Washington.