Zeller, D., R. Froese and D. Pauly. 2005. On losing and recovering fisheries and marine science data. *Marine Policy* 29: 69-73.

Available online at www.sciencedirect.com
SCIENCE DIRECT®
Marine Policy 29 (2005) 69–73
MARINE POLICY
www.elsevier.com/locate/marpol

On losing and recovering fisheries and marine science data

Dirk Zeller[a,*], Rainer Froese[b], Daniel Pauly[a]

[a] *Fisheries Centre, University of British Columbia, 2259 Lower Mall, Vancouver V6T 1Z4, Canada*
[b] *Institut für Meereskunde, Düsternbrooker Weg 20, 24105 Kiel, Germany*

Accepted 13 February 2004

Abstract

Large sums are spent annually collecting and, increasingly, electronically encoding field data, making them widely accessible. Earlier data were recorded on paper, and archived at a few institutions, which eventually discard them. Data recovery and distribution is a valuable contribution to science, as it counters the 'shifting baseline' syndrome and ensures long-term returns on funds society invested in data gathering. Data recovery need not be expensive. We present the data recovery from the Guinean Trawling Survey, conducted in the early 1960s off West Africa, which cost 0.2% of initial survey costs. Research and graduate training institutions, as well as funding agencies should make digital data globally available as part of their deliverables.

Keywords: Data loss; Data recovery; Fisheries; Surveys; Shifting baseline

1. Introduction

Globally, millions of taxpayers' money is spent annually to collect scientific data. In today's 'wired' world, most of these data are electronically encoded and are, theoretically, globally available (aided by web-based search engines). Examples can be found at the World Data Centres around the world (e.g., www.ngdc.noaa.gov/wdc/wdcmain.html#wdc). Alternatively, datasets are increasingly maintained by individual scientists or laboratories (e.g., the *Sea Around Us* Project database, http://www.seaaroundus.org; or R.A. Myers global stock recruitment database, http://fish.dal.ca/%7Emyers/welcome.html). This is a commendable development. However, only a few decades ago, computers were rare and data encoding an often difficult, expensive and highly specialized task [see Box 1 page 277 in Ref. [1] for discussion of problems related to the encoding of Indonesian trawl survey data]. Thus, until recently, most data were 'encoded' and stored on paper, limiting their distribution and availability. And unless the data were extensively analysed and published in the primary literature (usually in highly aggregated, summarized form), the underlying data and sampling design could not only be lost from scientific memory as it disappeared into the ever growing archives of libraries, but this loss might not even be perceived. Significantly, and often the norm for large scale fisheries surveys, was the archiving of paper copies of the data in a small number of institutions after the production of a limited number of official survey reports. This is a clear recipe for loss, not only from local, institutional memory, but also from the global scientific memory. Furthermore, libraries and institutions occasionally 'clean up' their holdings, often leading to loss of archived data; e.g., the German G.T.Z. (Deutsche Gesellschaft für Technische Zusammenarbeit; German Foundation for International Development) destroyed its archive of fisheries projects it supported in the past, including those in Indonesia documented in the book by Pauly and Martosubroto [1], and largely based on documents that they—fortunately—kept in their personal collections.

2. Why recover such data?

Data collection has a long history, and especially large-scale and long-term survey datasets collected decades ago provide a valuable function. They provide fundamental baselines of abundance, size structures and biodiversity patterns. These data can, among others be very useful for (1) contributions to time series data to be

*Corresponding author. Tel.: +1-604-822-1950; fax: +1-604-822-8934.
E-mail address: d.zeller@fisheries.ubc.ca (D. Zeller).

0308-597X/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.marpol.2004.02.003

used, e.g., for ecosystem modelling or improved stock assessments [2,3]; (2) establishment of biodiversity and geographic range indicators; or (3) the estimation of historic stock biomass and natural mortality rates of unexploited stocks. The last item is of great significance, as such unexploited stock estimates are very rare, yet highly valuable for understanding changes in exploited stocks over long time periods [4]. Thus, historic survey data sets are invaluable, inherent depositories of knowledge to counter the 'shifting baseline syndrome' [5].

It may be thought that recovering paper-based, archived data, and creating electronic databases thereof are expensive exercises. However, as we illustrate below, it is a much cheaper endeavour than the actual survey itself, which in most cases resulted in no more than the production of a limited number of hard-to-trace survey reports, the publication of a small number of papers [although often seminal ones, e.g., [6]], and the deposition of paper copies of data in host- and donor-country laboratories or library archives. The justification of and approach to such data recovery and their utility for retrospective analyses have been illustrated before [7]. Here, we present an example of a recovered dataset, and present estimates of costs for database creation, versus reported survey costs.

3. Example

Numerous large scale science and fisheries surveys were undertaken some 3–4 decades ago as part of the 'development' approach taken by developed countries during that period [7]. These included, among others a Gulf of Thailand Survey [8], an Indian Ocean Coast Survey [9], a North Pacific Survey [10], and the Guinean Trawling Survey in West Africa considered here [11].

The Guinean Trawling Survey (GTS) covered an extensive latitudinal gradient in the tropical Gulf of Guinea on the West African coast, from The Gambia in the North (12.1° N), to the mouth of the Congo river (5.5° S) in the South. The GTS was carried out in two phases corresponding to seasonal changes in oceanographic patterns in the Gulf of Guinea, with the first phase from August to December 1963, and the second phase from February to May 1964. Significantly, this period pre-dates the development of large-scale commercial fisheries for demersal resources in West African shelf waters.

The database created contains catch and length frequency data by taxon for all transect/station combinations sampled, for both day and night trawls. Data were taken from handwritten data sheets, which were kindly provided by Dr. Alan Longhurst on the request of the third author, and the senior author coordinated the creation of the database. Ancillary information, such as maps of bottom type, salinity, water temperature etc. is also included on the database CD-ROM. For ease of use and global availability we utilised Microsoft Access© software as database platform.

This database was first presented to the public at the international symposium on "*Marine fisheries, ecosystems and societies in West Africa: half a century of change*", held in Dakar, Senegal in June 2002. In particular, we believe that our colleagues in the West African countries should have ready access to this data and the underlying baseline ecosystem and fisheries knowledge contained therein. Thus, a large set of CD-ROMs containing the database and other ecosystem and regional information were produced and distributed to interested parties in West Africa and elsewhere [12].

Problems encountered during data recovery and database creation were mainly of a taxonomic nature. The taxonomic names encountered on the handwritten datasheets presented occasional difficulties, due to different field personnel identifying many species from (at that time) poorly known waters, a situation regularly encountered during surveys [13]. *FishBase* [www.fishbase.org, [14]] was used to update all taxa to the current status, and to confirm geographic ranges. In cases where the species name could not be determined (no matching synonym), the taxon was recorded as belonging to the next higher level (Genus or Family).

4. Cost comparison

The Guinean Trawling Survey of the early 1960s was conducted at a total cost of US$ 1,007,380 (1960s dollars), of which 72% was provided by the US Agency for International Development and 28% (in cash or kind) by other sources, including several Western European countries and the host countries in West Africa [11]. Adjusted to 2003 dollars, this translates into approximately US$ 17,000,000 for the entire GTS project. Cost comparison between survey costs (adjusted to 2003 US dollars) and database recovery costs (US$ 20,500), using the methods described in Torres et al. [13], indicates that the data recovery cost less than 0.2% of the initial survey costs (Table 1). Note that the costs presented here may be higher if commercial salary rates would be applied instead of academic salaries as was the case here. Nevertheless, even if costs were doubled, it would still represent less than 0.5% of original survey costs. Truly a 'cheap' deal.

5. Discussion

Datasets are the fuel that keeps the scientific engine running, and we should do more to ensure that our tanks remain full, especially with historic datasets. Unfortunately, there are many causes for data loss,

Table 1
Cost estimate for creating and encoding the Guinean Trawling Survey database, in 2003 US dollars. The 2003 cost base conversion for the total cost of the 1960s survey is indicated for comparison.

Person	Skill	Approx. cost US$ (year 2003)
Project leader	Conceptualising	1300
Scientist	Directing project, supervising, troubleshooting	6000
Research assistants	Data encoding, typing, verification	9000
IT specialist	Database design and integration	1200
Multimedia specialist	Incorporation of database and documentation onto distributable media	2000
	Approximate miscellaneous costs (copying, computing, etc.)	1000
	Total	20,500
	GTS survey cost	17,000,000

ranging from poor planning to major political and social disruptions [15]. The situation is exasperated by the common failure to appreciate that data that may seem uninteresting or unimportant to us today, may be a gold mine for future scientists [15–17]. A further cause for withholding, under-utilizing and eventually losing data is based on perceived ownership of data [15]. Many fisheries related data are (erroneously) perceived to be owned by entities other than the general public, despite marine resources being the 'common property and heritage of mankind' and surveys being funded by taxpayers [18]. Another area in which this ownership problem is large is in the academic arena of graduate student research and the associated data that are produced every year around the world. Many of these studies are never published in the peer-reviewed primary literature, and the associated data rarely get used beyond the requirements of the graduate degree. More significantly, often no-one besides the people associated with the studies (i.e., the student, supervisors and committee members) even know of the existence of these data. While more recently, research theses can be listed electronically (e.g., ProQuest Digital Dissertations, http://wwwlib.umi.com/dissertations/gateway), not all data-generating institutions subscribe to this service, nor are these works as readily available as peer review literature. Using interlibrary loans, graduate theses can be accessed, but the data themselves may not be readily available beyond a few years. Many institutions require data to be retained for a number of years, but often only in relation to published material. It appears that most academic institutions do not have policies to deal specifically with unpublished data. Many have general policies regarding data retention, but no concerted effort to make such data generally available after some protected period for priority publication by the authors. For example, James Cook University, Australia, the senior authors' *alma mater*, requires their schools to retain data for at least 5 years, and, while assigning priority right for publication to the authors of the work, makes no clear mention of the fate of data that is not published [19,20].

Other disciplines of science have established rules to deal with this (more or less officially codified). For example, in oceanography and taxonomy, unwritten professional codes allow about one year of withholding data or un-described taxa for priority publication by the original scientist before releasing the data or new taxa for description [15].

However, there are efforts underway to facilitate database availability, at least in some fields. For example, the *Global Biodiversity Information Facility* (GBIF, http://www.gbif.org) attempts to make global biodiversity data universally available.[1] Similarly, the *Oceans Biogeographic Information System* (OBIS, http://www.iobis.org) is a globally available, on-line node providing access to systematic, ecological, and environmental information systems about the ocean [21]. *FishBase* [www.fishbase.org, [14]] is a large database with key information on all fishes of the world. It is maintained by a team of specialists who scan through relevant scientific publications and extract and standardize data on, among other, population dynamics, reproduction, and trophic ecology of fishes. It also provides access to distributed online databases such as museum collections or trawl surveys, currently giving access to more than 1.5 million records in over 30 contributing databases. It also acts as an archive for historical data sets such as the GTS survey presented here, and it provides interfaces where users can upload survey or length-frequency data. *FishBase* receives nearly 10 million hits per month from a wide variety of users from all over the world, thus proving that there

[1] The following explanation is taken directly from the GBIF web page: "GBIF works cooperatively with and in support of several other international organizations concerned with biodiversity. These include the Clearing House Mechanism and the Global Taxonomic Initiative of the Convention on Biological Diversity, and regional biodiversity information networks. Functionally, GBIF encourages, coordinates and supports the development of worldwide capacity to access the vast amount of biodiversity data held in natural history museum collections, libraries and databanks. Participants in GBIF support network nodes through which they provide data. GBIF is evolving to be an interoperable network of biodiversity databases and information technology tools… In the long term, GBIF will provide a portal that enables simultaneous queries against biodiversity, molecular, genetic, ecological and ecosystem level databases, which will facilitate and enable "data mining" of unprecedented utility and scientific merit."

is public interest in scientific information if it is presented in a user-friendly fashion.

We suggest that policy developers and decision-makers push for the adoption of standard practise by all institutions (academic, governmental and otherwise) to implement temporal data restriction and release requirements (e.g., 5 years) to permit these agencies and scientists a first 'shot' at analysing and publishing the findings. However, thereafter all data should be made globally available and accessible, with assurance for permanent availability. Given the growing concerns about the effects of global issues such as climate change and overfishing on ecosystems and societies, large-scale, long-term datasets need to be freely available and accessible for future use and analysis [4,22,23]. Similar arguments have also been proposed with regards to global sharing of fisheries stock assessment data [2].

6. Conclusion

Data recovery, retention and distribution is a valuable contribution to science, as it:

1. assists in countering the 'shifting baseline' syndrome;
2. provides early time period anchor points for meta-analyses and modelling;
3. enables establishment of un- or little-exploited stock parameter values;
4. allows wider use and sharing of data by the scientific community; and
5. ensures longer term returns on the money society invested in initial data gathering.

The scientific community should endeavour to ensure that no data is 'lost' to future generations of scientists, decision makers and society in general. Digitally 'recovering' data that otherwise would only be available in printed archival versions ensures increased exposure and use of costly data and knowledge that otherwise are vulnerable to be 'forgotten' by future generations of potential users, and may thus become 'lost' in the ever growing archives and libraries of the world. Serious considerations should be given by academic and government institutions, as well as funding bodies to make digital data availability on a global scale part of their accountabilities and deliverables. After all, science is indeed forever [17].

Acknowledgements

We would like to thank the Pew Charitable Trusts, Philadelphia for funding the *Sea Around Us* Project, A. Longhurst for obtaining copies of the original archived datasheets, M.L.D. Palomares for taking the lead in creating the CD-ROM, S. Booth for careful data entry and quality check, C. Young for her multi-media skills structuring the CD-ROM, and the European Commission for funding its creation.

References

[1] Pauly D, Martosubroto P, editors. Baseline studies of biodiversity: the fish resources of Western Indonesia. ICLARM Stud. Rev. 23. Manila: International Center for Living Aquatic Resources Management; 1996. p. 321.
[2] Richards LJ, Schnute JT. A strategy for advancing stock assessment. In: Pitcher TJ, Hart PJB, Pauly D, editors. Reinventing fisheries management. London: Kluwer Academic Publishers; 1998. p. 399–406.
[3] Cox SP, Essington TE, Kitchell JF, Martell SJD, Walters C, Boggs CH, Kaplan I. Reconstructing ecosystem dynamics in the central Pacific Ocean, 1952–1998. II. A preliminary assessment of the trophic impacts of fishing and effects on tuna dynamics. Can J Fish Aquat Sci 2002;59(11):1736–47.
[4] Christensen V, Guénette S, Heymans JJ, Walters CJ, Watson R, Zeller D, Pauly D. Hundred-year decline of North Atlantic predatory fishes. Fish Fish 2003;4:1–24.
[5] Pauly D. Anecdotes and the shifting baseline syndrome of fisheries. Trends Ecol Evol 1995;10:430.
[6] Fager EW, Longhurst AR. Recurrent group analysis of species assemblages of demersal fish in the Gulf of Guinea. J Fish Res Board Can 1968;25(7):1405–21.
[7] Pauly D. Biodiversity and the retrospective analysis of demersal trawl surveys: a programmatic approach. In: Pauly D, Martosubroto P, editors. Baseline studies of biodiversity: the fish resources of Western Indonesia. Manila: ICLARM Studies Review 23; 1996. p. 1–6.
[8] Ritragsa S. Results of the studies on the status of demersal fish resources in the Gulf of Thailand from trawling surveys, 1963–1972. In: Tiews K, editor. Fisheries resources management in Southeast Asia. Berlin: German Foundation for International Development; 1976. p. 198–223.
[9] Hida T, Pereyra WT. Results of bottom trawling in Indian Seas by RV "Anton Bruun" in 1963. Proc Indo-Pacific Fish Counc 1966;11:156–71.
[10] Alverson DL, Pereyra WT. Demersal fish exploration in the northeastern Pacific Ocean—an evaluation of exploratory fishing methods and analytical approaches to stock size and yield forecasts. Proc Indo-Pacific Fish Counc 1972;13(3):224–54.
[11] Williams F. Report on the Guinean Trawling Survey. vols. I (601 p.), II (529 p.) and III (551 p.). Scientific, Technical and Research Commission, Organisation of African Unity: Lagos. 1968.
[12] Palomares MLD, Zeller D, Young C, Booth S, Pauly D. A CD-ROM on Northwest African marine ecosystems. November 2002. SIAP, SAUP, JRC and CRSP. Presented at the International Symposium on Marine Fisheries, Ecosystems and Societies in West Africa: Half a Century of Change, Dakar, Senegal; 24–28 June 2002. 2002.
[13] Torres F, Cabanban A, Bienvenida S, McManus J, Prein M, Pauly D. Using the NAN-SIS and FiSAT software to create a trawl survey database for Western Indonesia. In: Pauly D, Martosubroto P, editors. Baseline studies of biodiversity: the fish resources of Western Indonesia. Manila: ICLARM Stud. Rev. 23; 1996. p. 276–83.
[14] Froese R, Pauly D, editors. FishBase 2000: concepts, design and data sources. Los Baños, Laguna, Philippines: ICLARM (updates available at www.fishbase.org); 2000. p. 344.
[15] Mathews CP. On preservation of data. NAGA (ICLARM) 1993;16(2–3):39–41.
[16] Pauly D. Data-rich books. BioScience 1992;43(3):167–8.
[17] Janzen D. Science is forever. Oikos 1986;46:281–3.

[18] Russ GR, Zeller D. From mare liberum to mare reservarum. Marine Policy 2003;27(1):75–8.
[19] Anonymous. Handbook for research higher degree students. James Cook University, Australia; 2000. http://www.jcu.edu.au/courses/hanbooks/research/pghandbook.html.
[20] Anonymous. Statement and guidelines on research practice. James Cook University, Australia; 2000. http://www.jcu.edu.au/office/policy/resresp.htm.
[21] Stocks KI, Zhang Y, Flanders C, Grassle JF. OBIS: Ocean biogeographic information system, www.iobis.org. The Institute of Marine and Coastal Science, Rutgers University. 2000.
[22] Watson R, Pauly D. Systematic distortions in world fisheries catch trends. Nature 2001;414:534–6.
[23] Pauly D, Christensen V, Guénette S, Pitcher TJ, Sumaila UR, Walters CJ, Watson R, Zeller D. Towards sustainability in world fisheries. Nature 2002;418:689–95.