

OPEN

DATA DESCRIPTOR

Biodiversity data rescue in the framework of a long-term Kenya-Belgium cooperation in marine sciences

Carolien Knockaert¹, Lennert Tyberghein¹, Annelies Goffin¹, Delphine Vanhaecke¹, Harrison Ong'anda², Enock O. Wakwabi² & Jan Mees¹

The Kenya-Belgium data collection includes about 111,800 biotic observations on benthos, algae, fish, zooplankton, phytoplankton, birds and mangroves which cover more than 400 unique locations that were sampled between 1873 and 1999. The scope of this data digitization project was to recover data in theses and reports resulting from marine and coastal research activities in the Eastern African region conducted between 1984 and 1999. Data were digitized and quality checked in the frame of the Belgian LifeWatch project. The dataset provides a better insight into the different types of research conducted between 1985 and 1996 in frame of the Kenya-Belgium cooperation in marine sciences (KBP) project and can facilitate further coastal biodiversity research in Kenya.

Background & Summary

The Kenya-Belgium cooperation in marine sciences (KBP) project was launched in 1985¹ as an initial collaboration between the Free University of Brussels (VUB) and the Kenya Marine and Fisheries Research Institute (KMFRI) under the supervision of the late Prof. Dr. Philip Polk. The aim was to improve the collaboration between different scientific institutes and marine scientists in Belgium and Kenya. The project consisted of three successive phases. The first phase of the KBP, named Cooperation in the field of marine ecology and management of the coastal zone (1985–1988) included an update of the necessary sampling equipment (basic lab, cars, boats and computers) and the start of a postgraduate course Fundamental and Applied Marine Ecology (FAME) at VUB for scientists of developing countries. The first fundamental research efforts were logically directed at the inventory and description of fauna and flora at the Kenyan Coast (Gazi Bay). The initial phase was extended for another four years (1989–1992) in a new project called Higher institute for marine sciences. The main objective was to conduct fundamental research in the areas of plankton, reef ecology, water chemistry, coastal oceanography and modelling, fisheries, algae, pollution and library sources. The successful cooperation between Kenya and Belgium and the coordinating role of the KBP proved to be a major attraction to other countries in East-Africa and Europe and resulted in several major international projects. As a logical extension, a third phase of the KBP started with a project on Research towards sustainable exploitation of natural resources in mangrove forests (1992–1996) with the continuation of the research on fish, crustaceans, birds, algal and phytoplankton cultivation and pollution monitoring. Most important realizations in this phase were the development of African's largest oyster farm in Gazi Bay and the reforestation of more than 10 hectares of mangrove forests (90,000 trees of 5 different mangrove species). Between 1985 and 1996, a total of 13 KMFRI researchers completed the FAME course at the VUB. During this period, the number of scientific publications, especially in international journals, increased considerably (Table 1). In 2012, the long lasting collaboration as described above resulted in a formal Memorandum of Understanding (MoU) between the Flanders Marine Institute (VLIZ) and the Kenya Marine and Fisheries Research Institute (KMFRI). This MoU aims to promote further partnership in the field of marine sciences between Belgium and Kenya (in a coordinated way). One of the activities within this MoU is the recovery of data resulting from marine research in Kenya. In the past, the data collected during the KBP were

¹Flanders Marine Institute, Wandelaarkaai 7, 8400, Oostende, Belgium. ²Kenya Marine and Fisheries Research Institute Mombasa Centre, P.O. Box 81651, 80100, Mombasa, Kenya. Correspondence and requests for materials should be addressed to C.K. (email: carolien.knockaert@vliz.be) or L.T. (email: lennert.tyberghein@vliz.be)

Source	1975–1979	1980–1984	1985–1989	1990–1994
Theses	0	9	8	9
Local journals and local symposia	3	24	30	22
International journals	1	4	27	34
Proceedings intern. symp.	0	5	3	3
Total	4	42	68	68

Table 1. Scientific publications by KMFRI researchers between 1975 and 1994. Adapted from Seys, J.¹.

mostly scattered and even hidden in the publications and reports coming out of the project. In that way, such data are nowadays hard to find for researchers: a lot of time may be lost when figuring out either where to find the data in literature or what type of data exist in a particular scientific work. Thanks to this project potentially high interesting data were digitized, standardized, quality controlled and brought together in one comprehensive dataset. Overall, this data rescue project focused on the digitization and online disclosure of distribution records of observed taxa in both time and space to make these data accessible again for the scientific community. The Kenya-Belgium data collection is online available through the LifeWatch portal (<http://www.lifewatch.be/en/marine-data-archeology>) and can be assessed through the Integrated Marine Information System (IMIS) hosted at VLIZ.

Methods

Data inventory. The first step was to identify potential literature for digitization. The Belgian Marine Bibliography or BMB (<http://www.vliz.be/en/belgian-marine-bibliography>) was searched to set up a Kenyan literature collection. The BMB is a reference list of publications on the Flemish coast and the Belgian Part of the North Sea as well as other marine, estuarine and coastal publications written by Belgian authors and scientists or foreign scientists affiliated with Belgian institutes. The BMB query resulted in a total of 402 theses and reports. First authors and affiliated institutes were contacted to gain approval and to comment on the digitization of the data in each publication. After setting priorities (selection of all publications concerning marine research in Kenya and the Eastern African region and approval by original authors and/or Belgian institutes for digitization and disclosure of the data), a total of 70 publications (29 reports and 41 theses) were selected for data digitizing (Online-only Table 1).

Quality control, standardization & data output. Data were digitized using a data format containing standard fields for a correct understanding of the data. The most important fields are the location linked through a Well-Known Text (WKT), sampling protocol and sample size, event date, scientific name, sex, lifestage and related measured biotic (describing the living organisms in the ecosystem) or abiotic parameters (chemical or physical parameters describing the environment of the observed species). Metadata were added by searching the original publications. The species taxonomy was verified using the taxon match tool offered by the World Register of Marine Species (WoRMS - <http://www.marinespecies.org>), an authoritative and comprehensive list of names of marine organisms edited and reviewed by an international team of more than 240 taxonomic editors world-wide. Every species has a unique identifier² known as the AphiaID. This identifier links the species name to an internationally accepted standardized name and associated taxonomic information, and also redirects to the most accurate information on the species taxonomy, (e.g. accepted names and synonyms). Next to the taxonomic information basic trait information was added in the database using Marine Species Traits (<http://www.marinespecies.org/traits>). If not present, geographical coordinates were added to the sampling locations either by georeferencing in QGIS based on the maps present in the particular studies or by matching with the Marine Regions Gazetteer³. In the next step datasets were created and described in the Integrated Marine Information System (IMIS) and archived in the Marine Data Archive (MDA - <http://mda.vliz.be>). In the last step a Digital Object Identifier (DOI) was assessed to each dataset making them more visible, traceable and citable. A total of 86 datasets were finished and assigned with a DOI. Biotic datasets were transferred and integrated in AfroBIS (<http://afrobis.csir.co.za>), the African node of the Ocean Biogeographic Information System (OBIS) and can be downloaded through the EurOBIS (<http://www.eurobis.org>)/EMODnet Biology (<http://www.emodnet-biology.eu>) platform and via the KMFRI IPT (<http://ipt.vliz.be/kmfri>) hosted at VLIZ.

Data selection. The Kenya-Belgium data collection focused on the temporal and geographical distribution of biotic data. Experimental data (e.g. cage exclusion experiments) and data on stomach analysis of fish species were excluded from the collection. Eventually 68 unique biota datasets were selected to integrate in the data collection (Online-only Table 2).

Data Records

The Kenya-Belgium data collection contains a total of 111,784 biotic observations in the Eastern African region dating from 1873 till 1999.

Geographical coverage. Samples were collected at 408 point locations (Fig. 1). Most of these were located along the Kenyan Coast, more specifically at Gazi Bay, and represented about 87% of database records. This is easily explained because the first research efforts made in frame of the KBP were to make an inventory and description of fauna and flora at the Kenyan Coast and Gazi Bay. The second largest collection of observations originated

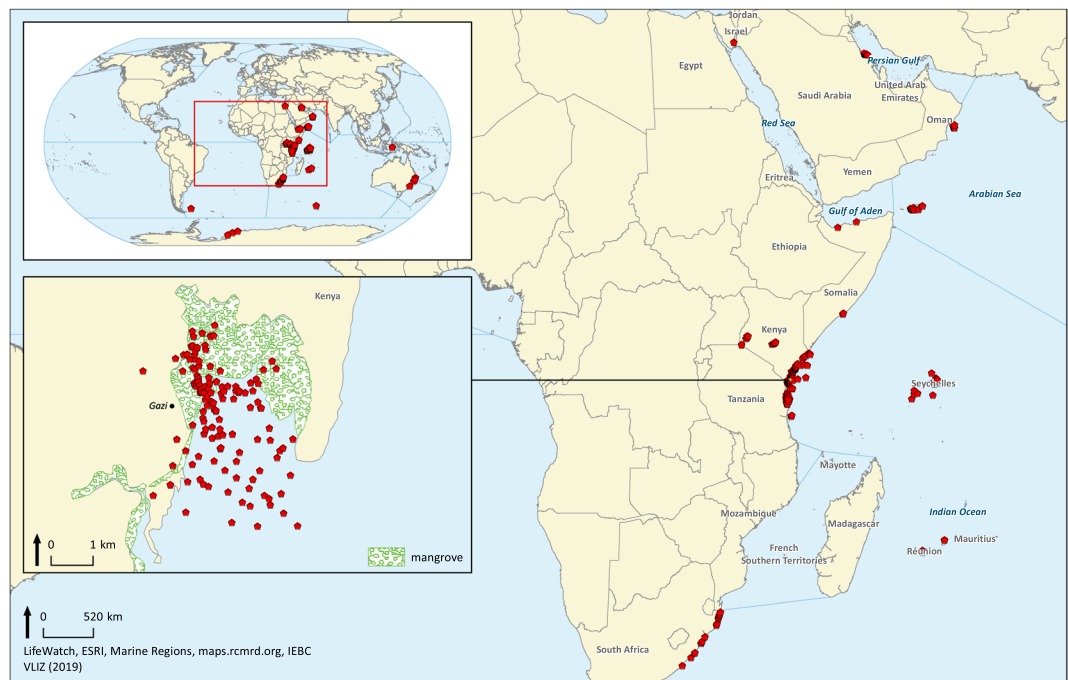


Fig. 1 Georeferenced map of all biological observations and sampling locations at Gazi Bay, Kenya.

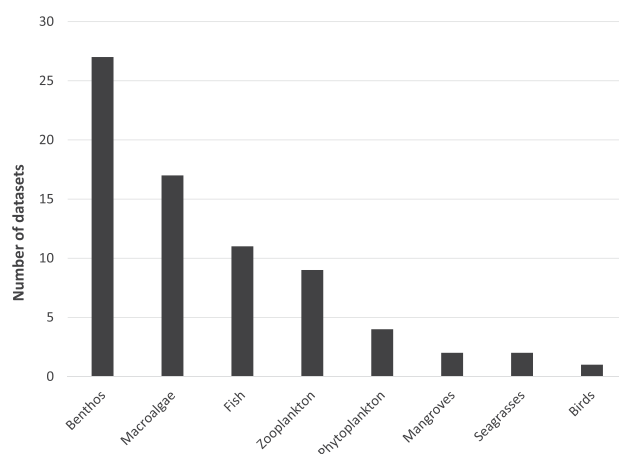


Fig. 2 Thematic coverage of datasets.

from the coastal area of Tanzania, followed by the East Coast of South Africa. Furthermore, a substantial part of the dataset observations came from the Seychelles and Socotra.

Themes & temporal coverage. The biological datasets were classified into 8 groups based on the number of observations belonging to a specific functional group (lower limit 70%) (Fig. 2). Data are available from 1873 to 1999 (Fig. 3). Data records from 1873, 1875 and 1964 are literature records and were not part of research field surveys. The increase in number of datasets per year is clearly correlated with the launch of the KBP and the start of the FAME course for KMFRI researchers. The number of field surveys increased immediately in 1985 and decreased again in 1997 (end of successive phase 3 of the KBP). The diversity of research topics was the highest in 1991 and 1992 and could be linked to KBP phase 2 for which the main objective was to conduct multidisciplinary research on plankton, reef ecology, fisheries, algae and dynamics of mangrove ecosystems.

Taxonomic coverage. 99.8 % of original scientific names have been referenced through the WoRMS database representing in total 2,198 unique aphiasIDs. Within these unique aphiasIDs 78.8% of used names were accepted in literature, 18.2% was not accepted in literature, 1.5% were alternate representations (species that are represented twice: once with and once without subgenus) and 1.3% had an uncertain status (unassessed). 67.5% were identified to the species level (1490 unique aphiasIDs). 750 records were not suitable (group names e.g. meiobenthos, non-identified, eggs, zooplankton) for a taxonomic match in WoRMS. These records together with non-matches (0.2%) were not discarded. The unmatched names may be added to the register after approval

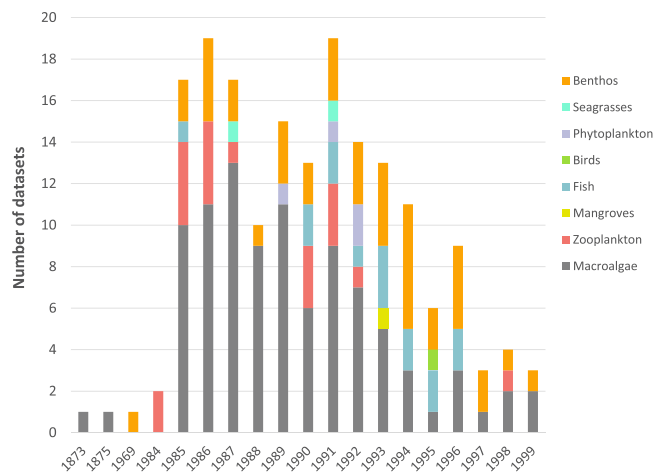


Fig. 3 Temporal distribution of biological datasets per theme.

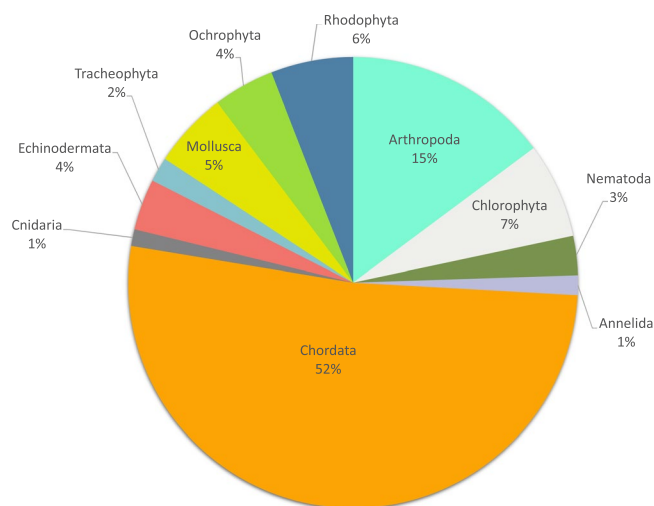


Fig. 4 Taxonomic coverage of data collection: percentage of observations per phylum.

of the dedicated editor. The total Kenya-Belgium data collection includes records for 29 different phyla (Fig. 4). Half of the observation records belonged to the Chordata (represented by birds and fish taxa) and although this frequency is almost 12 times higher than the Ochrophyta observations (4.3%), the number of unique taxa for both phyla were almost the same (Fig. 5). Remarkable are the 142 observations on Myxozoa listed in the database, as these represent 53 unique taxa. An important research topic within the Kenya-Belgium Project (KBP) was the description and inventory of benthic fauna in the mangroves of Gazi Bay. 15.8% of benthic invertebrate observations in Gazi Bay were identified at species level. The benthic invertebrate species community belonged to only 3 phyla and 6 classes, of which the Mollusca were most represented (57.7%). 30% of mollusc observations were on the mangrove oyster *Crassostrea cucullata*, and more specifically on its ecomorphology (physical properties of the shell and live biomass of observed species). Within the phylum of Arthropoda 2 classes were identified: Hexanauplia, represented by the benthic copepods (20 unique species), and Malacostraca of which all taxa belonged to the order of Decapoda, and more specifically to the Brachyuran mangrove crabs (56 unique species).

Documentation and dataset dissemination. The Kenya-Belgium data collection⁴ is accessible through IMIS and can be downloaded from the Marine Data Archive (MDA).

Technical Validation

The Kenya-Belgium database is a compilation of 68 biotic datasets which were digitized in the frame of a data rescue project. Focus was the digitization of historical biological data sampled in the Eastern African Region and/or in the frame of the Kenya-Belgium cooperation in marine sciences project (KBP) or collected by Belgian or Kenyan marine researchers. Many sources of historically collected data exist, but are scattered in different databases, hidden in papers or reports or even stored on old storage media such as cassettes, disks, tapes, cd's, etc. There is an enormous amount of information already collected and stored in a standardized way about the world's biodiversity. To date most of this old information has not been digitized. The digitization strategy used in

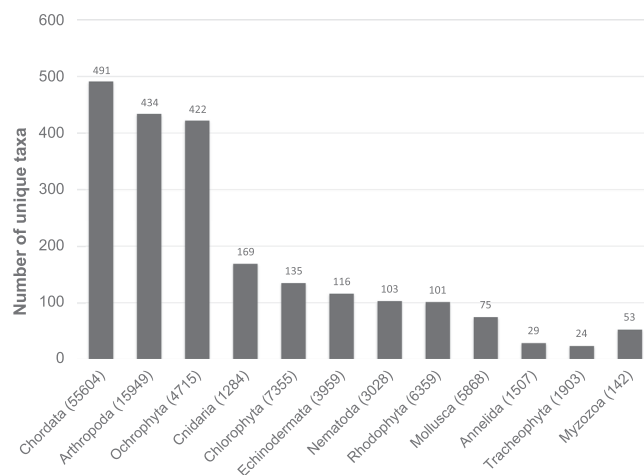


Fig. 5 Taxonomic coverage of data collection: number of unique taxa per phylum.

this work was to stay as close to the original source data as possible. However, data from different sources always needs to be interpreted with caution. Hence, users of this dataset need to consider two caveats. First, not all data records hold data for all parameters/variables present in the final dataset, simply because they were not measured. Therefore, it might seem that there are quite some gaps. However, on the contrary, providing this information as well only increases the completeness of this dataset. Second, despite the effort to search the original publications for missing metadata, there are obviously still gaps (e.g. sampling date -and location) due to the fact that they were not properly described in the original publications. Notwithstanding the above caveats, the digitizing of historical data is crucial to fill in the spatial and temporal gaps in the data that is currently available to science. The current availability of biodiversity data from this region is still too low to give an adequate insight in the underlying processes that control the functioning of our ecosystems. This data rescue project certainly has improved and contributed to create a better accessibility (open access) and visibility of the data to the scientific community. The species list will be used to contribute to AfReMaS, the African Register of Marine Species - <http://www.marinespecies.org/afremas>), a taxonomic database of marine species found along the African coast. Next to that, these revitalized datasets can serve as a starting point for the biodiversity work in the Second International Indian Ocean Expedition (<http://www.iioe-2.incois.gov.in>) program. The sustainable use and management of biodiversity will require data about it to be available when and where decision-makers and scientists alike need that type of information.

Usage Notes

All data are publicly available in Open Access and can be used under the license of CC BY with acknowledgement of the authors.

References

1. Seys, J. *The Kenya-Belgium cooperation in marine sciences: a review of ten years scientific collaboration*. (VLIZ, 1995).
2. Costello, M. J. *et al.* Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. *PLoS ONE* **8**, e51629 (2013).
3. Claus, S. *et al.* Marine Regions: towards a global standard for georeferenced marine names and boundaries. *Mar. Geod* **37**, 99–125 (2014).
4. Knockaert, C. *et al.* Historical Kenyan-Belgian research datasets (1873–1999). *Marine Data Archive*, <https://doi.org/10.14284/331> (2019).

Acknowledgements

This work makes use of data and/or infrastructure provided by VLIZ and funded by Research Foundation - Flanders (FWO) as part of the Belgian contribution to LifeWatch (FWO project I000819N). LifeWatch was established as part of the European Strategy Forum on Research Infrastructure (ESFRI) and can be seen as a virtual laboratory for biodiversity research. Special thanks to the KMFRI staff providing data and information to fill in the gaps concerning the Kenyan datasets. Thanks to Nathalie De Hauwere and Britt Lonneville for GIS support.

Author Contributions

Carolien Knockaert: digitization and quality control of data, dataset descriptions and DOI requests, compilation of data, drafting and revising the article. Dr. Delphine Vanhaecke: acquisition of data and initial quality control, revision of the article. Annelies Goffin: acquisition of data and revising of the article. Dr. Lennert Tyberghein: drafting and revising the article.

Additional Information

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019