GCRMN
Global Coral Reef Monitoring Network

ICRI
International Coral Reef Initiative

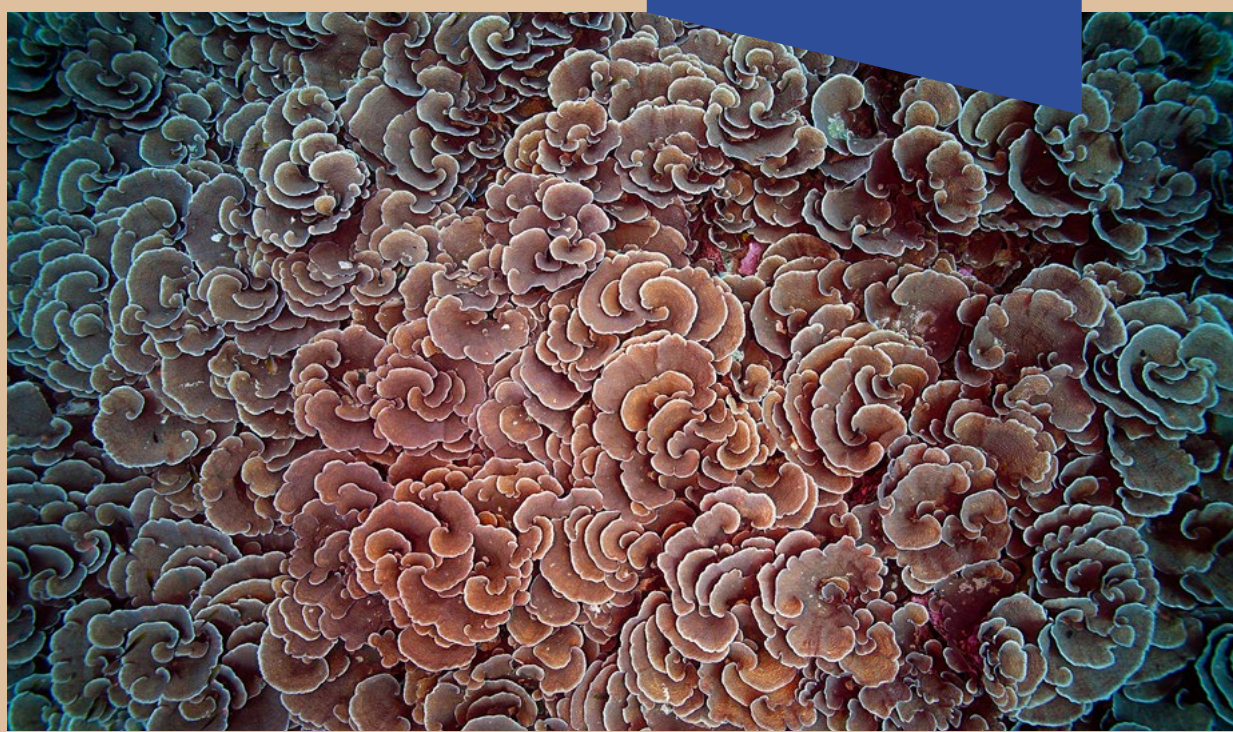
Australian Government

Australian Institute of Marine Science

Status of Coral Reefs of the World: 2020

Chapter 13. Data collation and processing

Edited by: David Souter, Serge Planes, Jérémy Wicquart, Murray Logan, David Obura and Francis Staub



Australian Government
Department of Foreign Affairs and Trade

Gouvernement Princier
Principauté de Monaco

Government Offices of Sweden
Ministry of the Environment and Energy

Prince Albert II of Monaco Foundation

UN environment programme 1972-2022

The conclusions and recommendations of this report are solely the opinions of the authors, contributors and editors and do not constitute a statement of policy, decision, or position on behalf of the participating organizations, including those represented on the cover.

Chapter 13.

Data collation and processing

Author: Jérémy Wicquart

1. Data acquisition

Based on advice from GCRMN regional coordinators and with the support of the International Coral Reef Initiative (ICRI), owners and custodians of data previously provided to the GCRMN, regional organizations, NGOs and researchers were approached to contribute coral reef monitoring data to the GCRMN *Status of Coral Reef of the World: 2020* report. Data sharing agreements were signed with each data contributor, which governed how their data could be used and provided assurances that their contribution would be recognised appropriately in associated GCRMN outputs. Only raw data for which the contributors were considered official custodians were collated. Except in rare cases, data extracted from scientific literature were not included because these data often lack complete metadata. Where necessary, data were homogenized in consultation with data contributors in order to maximize the reliability of the final results. Data acquisition was conducted throughout 2019 and required 12 months to complete. As a consequence, the majority of data on which the report was founded pre-date 2019.

2. Data homogenization and processing

Numerous monitoring programs have been established around the world at different times, for different purposes and using different protocols. Some methodological standards (e.g. GCRMN[1], Atlantic and Gulf Rapid Reef Assessment (AGRRA), Reef Check) have emerged during the last two decades but different standards tend to be used in different regions and by different monitoring programs. Thus, datasets collected by different coral reef monitoring programs differ in their formats, and use different variables, units and taxonomic resolution. As a consequence, it was essential to implement a rigorous process to standardize the format of all contributed datasets in order to create a unique and homogenous global dataset for quantitative analysis. All data homogenization was performed by a single person within the data analysis team in order to ensure consistency, provide a single point for issue tracking and reduce the burden on data contributors (Fig. 13.1).

[1] English, S.A, Wilkinson, C., Baker, V.J. (1997). Survey manual for tropical marine resources. 2nd Edition. Australian Institute of Marine Science, Townsville, Australia. 378p.

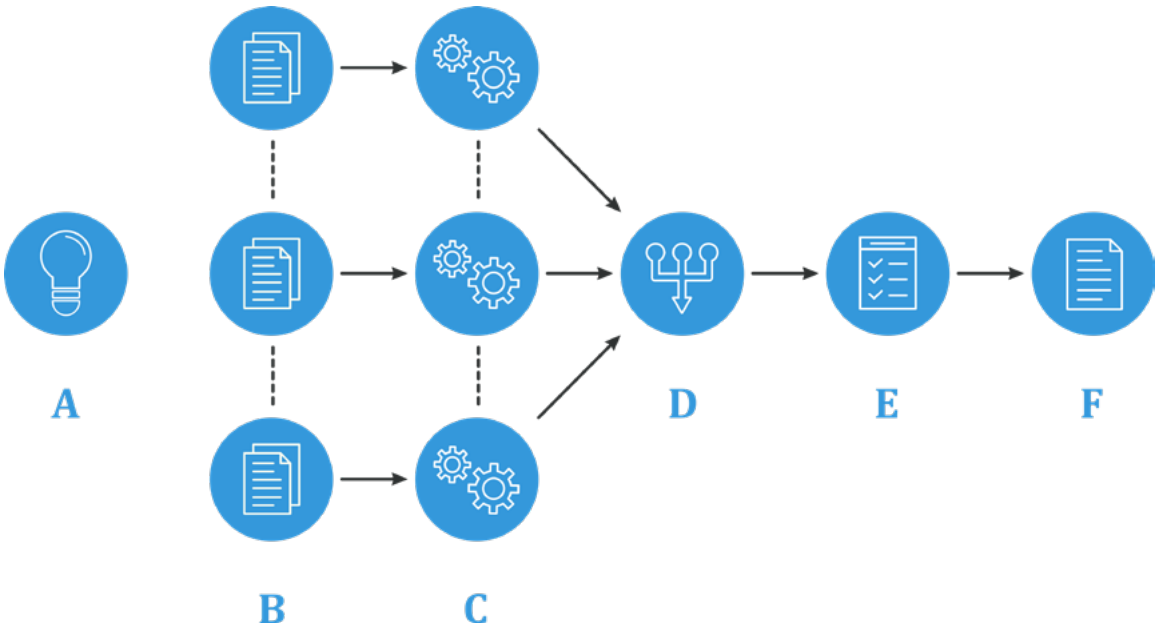


Figure 13.1. Steps used in the data homogenization and cleaning process. A: selection of variables and levels; B: export of individual raw datasets in csv format (i.e. as provided by data contributors); C: cleaning of all datasets individually; D: merging of all individually cleaned datasets; E: quality assurance and quality control (QAQC); F: exportation of the global dataset.

The first step in the homogenization process was to define the variables required for the synthetic global dataset (Step A, Fig. 13.1). The 22 variables listed in tab. 13.1 were selected. These variables spanned four broad groups: spatial variables (2 to 12), temporal variables (13 and 14), methodological variables (15 and 16) and taxonomic variables (17 to 21). Spatial and taxonomic variables were nested. For example, a given location can include several sites, each of which could be comprised of several replicates.

Table 13.1. Variables included in the synthetic global dataset.

	Variable	Type	Description
1	DatasetID	Factor	Dataset ID
2	Area	Factor	GCRMN region (see Fig. 13.2)
3	Country	Factor	Country
4	Archipelago	Factor	Archipelago
5	Location	Factor	Location or island
6	Site	Factor	Site within the location
7	Replicate	Integer	Replicate ID
8	Quadrat	Integer	Quadrat ID
9	Zone	Factor	Reef zone
10	Latitude	Numeric	Latitude of the site
11	Longitude	Numeric	Longitude of the site
12	Depth	Numeric	Mean depth at which data were collected
13	Year	Integer	Year in which data were collected

14	Date	Date	Date (YYYY-MM-DD) on which data were collected
15	Method	Factor	Method used to collect the data
16	Observer	Factor	Name of individual who collected the data
17	Category	Factor	See Tab. 13.2
18	Group	Factor	See Tab. 13.2
19	Family	Factor	Family name
20	Genus	Factor	Genus name
21	Species	Factor	Species name
22	Cover	Numeric	Percentage cover

Next, raw datasets were converted into csv format (Step B, Fig. 13.1) and then individually homogenized (Step C, Fig. 13.1). Homogenization consisted of:

1. Deleting, renaming and adding variables (to be consistent with those listed in Tab. 13.1); and
2. Ensuring consistency in the format of latitude and longitude (e.g. from hexadecimal to decimal format), date (e.g. from DD-MM-YYYY to YYYY-MM-DD), and the units for depth (e.g. from feet to meters) and cover (e.g. number of points counted on a transect to percentage cover).

The positions of sites were visually verified using an interactive map. When data were missing or ambiguous, clarification was sought from data contributors.

Standardized datasets were then merged (Step D, Fig. 13.1). In order to deal with the variation in the taxonomic level at which benthic data were recorded by different monitoring programs, it was necessary to standardize records at an equivalent level. This was achieved by ensuring that each record was completely described by five variables (*Category, Group, Family, Genus and Species*). The variables *Category* and *Group* (Tab. 13.2) were adapted from English et al. (1997)[1]. The variables *Family, Genus* and *Species* reflect actual taxonomic levels and their validity was assessed using *World Register of Marine Species* (WoRMS)[2]. Particular attention was given to genus names that were identical between distinct taxonomic groups to avoid re-categorization errors. For example, *Turbinaria* is a genus of both algae and coral.

[2] WoRMS Editorial Board (2022). World Register of Marine Species. Available from https://www.marinespecies.org at VLIZ. Accessed 2021-11-03. doi:10.14284/170

Table 13.2. Selected levels for the variables Category and Group.

Category	Group
Abiotic	Rock
	Rubble
	Sand
	Silt
Algae	Coralline algae
	Macroalgae
	Turf algae
Hard bleached coral	
Hard dead coral	
Hard living coral	
Other fauna	Actiniaria
	Alcyonacea
	Antipatharia
	Asteroidea
	Bivalvia
	Bryozoa
	Corallimorpharia
	Crinoidea
	Decapoda
	Echinoidea
	Gastropoda
	Holothuroidea
	Hydrozoa
	Ophiuroidea
	Polychaeta
	Porifera
	Tunicata
	Zoantharia
Seagrass	

The final step in the data homogenization process was quality assurance and quality control (QA/QC) (Step E, Fig. 13.1). This was achieved by first calculating the sum of percentage covers of all categories at the lowest sampling unit (e.g. transect). The natural assumption is that the sum of all percentage covers would equal 100%. However, this was not always the case. As a consequence, the following QA/QC protocols were applied based on the sum of the percent cover calculated for each sample:

1. Percent cover lower than 0% - This result was possible only when there was an error either in data entry by a data contributor or an error in data homogenization. After verification of the data cleaning process for the corresponding dataset (Step C, Fig. 13.1) during which corrections were made if needed, all samples with total cover lower than 0% were removed from the global dataset.

2. Percent cover between 0% and 100% - This occurred when observations of some cover categories (e.g. non-living substrates, tape, wand, shadows) were removed from a sample by data contributors or if data were collected on only a specific subset of benthic cover categories (e.g. living hard living coral). This was acceptable and all corresponding samples were retained in the global dataset.

3. Percent cover equal to 100% - This was the best case and occurred when all information in a sample was available. All corresponding samples were retained in the global dataset.

4. Greater than 100% - This result was possible only when there was an error either in data entry by a data contributor or an error in data homogenization. In this scenario, the data cleaning process for the corresponding dataset was verified (Step C, Fig 13.1) and corrections were applied if necessary. If the data cleaning process was accurate, further investigation was conducted. Occasionally, the total cover was very close to but still exceeded 100%. This occurred when the data provided were rounded averages rather than raw data. In order not to exclude these potentially valid data, a threshold of 101% was applied and percent covers within such samples were reduced to achieve a total cover of 100%. If, after verification of data cleaning and the application of corrections, the sum of percent covers within a sample remained greater than 100%, the sample was removed from the global dataset.

All data homogenization procedures were done within the R Statistical and Graphical Environment (version 3.6.3) mainly using packages contained in the tidyverse (version 1.3.1)[3].

3. Limitations

The data homogenization process was designed to eliminate a maximum number of errors, which sometimes led to a significant loss of data (up to 10% of samples within a given dataset), usually during the QA/QC step (Step E, Fig. 13.1). Due to the great diversity of categories used by data contributors, it was not possible to implement an automatic re-categorization process. This step was conducted manually and, as a consequence, may have introduced errors. In order to reduce the influence of potential errors introduced in key categories (e.g. living hard coral), individual trends were compared with those reported in associated documents (e.g. reports, scientific articles) provided by data contributors or, when uncertainty remained, clarification was sought from the data providers.

Description of homogenized data

The data homogenization process made it possible to build a global dataset based on the aggregation of data contained in 248 datasets, collected from 12,160 monitoring sites and provided by more than 300 contributors.

All data were assigned to one of the 10 GCRMN regions (Fig. 13.2) for analysis and reporting. The boundary of each region broadly corresponded with historical GCRMN regional boundaries based on existing national or informal networks.

The total area of coral reefs within each GCRMN region varies greatly, ranging from 780 km² in the Eastern Tropical Pacific to 78,272 km² in the East Asian Seas region, which includes the Coral Triangle (Tab. 13.3). The East Asian Seas, Pacific and Australia regions together account for almost 73% of world's coral reef area.

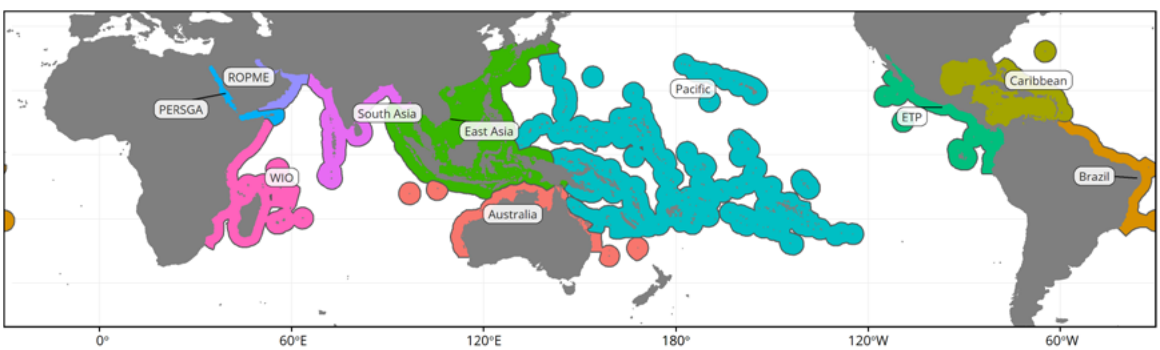


Figure 13.2. The 10 GCRMN regions. ETP is the Eastern Tropical Pacific. PERSGA is the area included within the Regional Organization for the Conservation of the Environment of the Red Sea and Gulf of Aden. ROPME is the sea area surrounded by the eight Member States of the Regional Organisation for the Protection of the Marine Environment. WIO is the Western Indian Ocean.

[3] Wickham, H. et al., (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Data were contributed from all 10 GCRMN regions (Fig. 13.3A). Eighty percent of sites surveyed were located in the Pacific, East Asian Seas and Caribbean regions. The patchiness and remoteness of reefs in some regions limited the spatial coverage of surveys, particularly in the Pacific and East Asian Seas regions which have the greatest areas of coral reefs. The vast majority of surveys were conducted at depths shallower than 20 m, with 25% conducted at 5 m (Fig. 13.3D).

Table 13.3. Summary statistics for each GCRMN region describing the area of coral reefs and the number sites and long-term monitoring sites from which data were compiled for the global dataset. A site is a unique GPS position where data were recorded. A site was considered a long-term monitoring site if the time between the first survey and the most recent survey was greater than 15 years, and may have been surveyed multiple times in the interim.

Region	Reef area*		Sites		Long term monitoring sites	
	Reef area (km^2)	Proportion of total reef area	Total Number	Proportion of global dataset	Total Number	Proportion of global dataset
East Asian Seas	78,272	30.15	2,570	21.13	158	26.87
Pacific	69,424	26.73	4,050	33.31	50	8.5
Australia	41,802	16.1	372	3.06	157	26.7
Caribbean	26,397	10.17	3,166	26.04	135	22.96
Western Indian Ocean	15,179	5.85	915	7.52	64	10.88
Red Sea and Gulf of Aden	13,605	5.24	243	2	7	0.01
South Asia	10,949	4.22	389	3.2	9	1.53
ROPME Sea Area	2,009	0.77	68	0.56	0	0
Brazil	1,226	0.47	35	0.29	9	1.53
Eastern Tropical Pacific	780	0.3	352	2.89	6	1.02

*** World Resources Institute. Tropical Coral Reefs of the World (500-m resolution grid), 2011. Global Coral Reefs composite dataset compiled from multiple sources for use in the Reefs at Risk Revisited project incorporating products from the Millennium Coral Reef Mapping Project prepared by IMaRS/USF and IRD. https://datasets.wri.org/dataset/tropical-coral-reefs-of-the-world-500-m-resolution-grid**

Most surveys were conducted after 2005, with the proportion of surveys conducted in each year increasingly rapidly until 2016/17 (Fig. 13.3C). The decline in the proportion of surveys conducted after 2017 was likely an artefact of the timing of the data acquisition and collation process which occurred during 2019 and thus potentially before contributors had fully collated or published their most recent survey data.

More than 75% of sites were surveyed only once (Fig. 13.3B). The high proportion of single surveys was attributable to the widespread adoption random sampling designs that were based on surveys of haphazardly chosen sites that are unlikely to be revisited.

Repeated surveys of fixed sites were conducted by some monitoring programs, although the time span over which sites were monitored was generally less than 10 years (Fig. 13.3B). Only 2% of sites were considered long-term monitoring sites, with data collected over periods greater than 15 years. The greatest proportion of long-term monitoring sites occurred in the East Asian Seas, Australia and Caribbean regions (Tab. 13.3, Fig. 13.3A).

The use of fixed or random sites has profound implications for data analyses and interpretation of results. Random sampling typically provides less biased estimates of reef condition and potentially better spatial coverage, whereas repeated surveys of fixed sites provide greater power to detect change and more precise estimates of temporal trends.

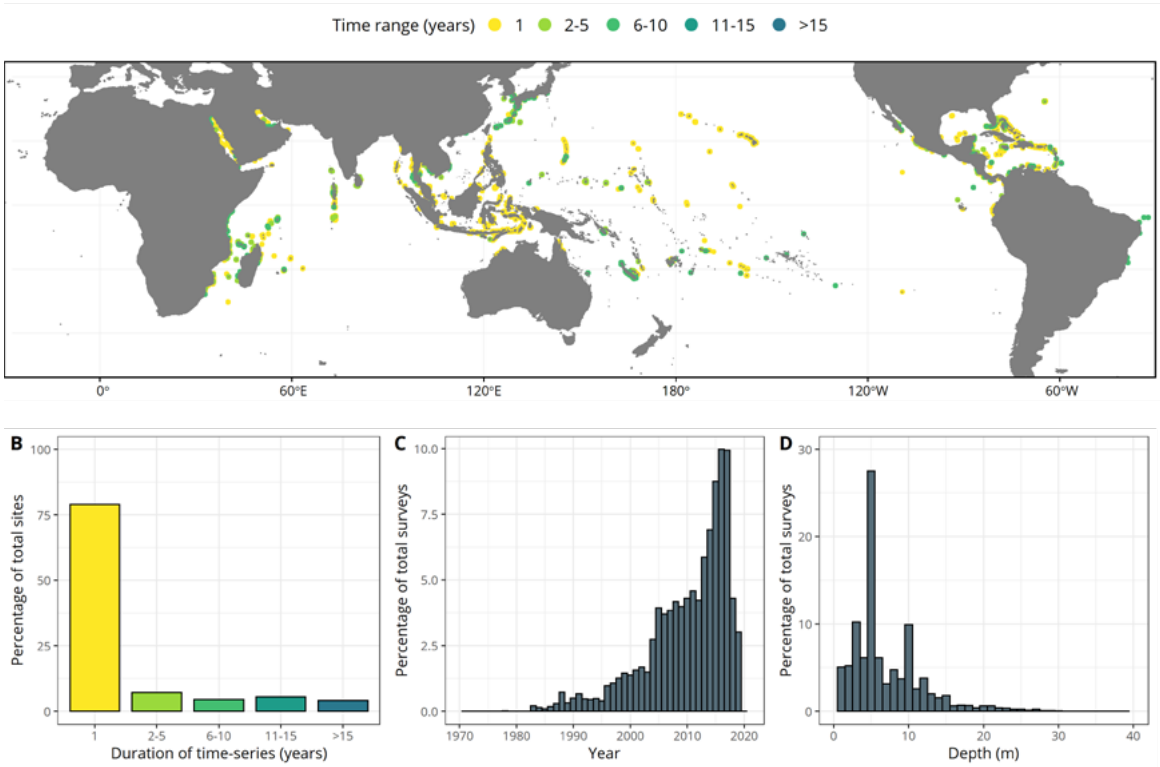


Figure 13.3. Distribution and duration of monitoring sites across the world (**A**), proportion of sites within each category describing the time span between the first and most recent surveys (**B**), proportion of the total number of surveys conducted in each year (**C**) and percentage of the total number of surveys by depth (**D**). For figures 13.3A and B, colours represent the time span between the first survey and the most recent survey at each site.

CORDIO

CRIOBE
USR3278
Centre de Recherches Insulaires et Observatoire de l'Environnement