DATA EXPLORATION & PREPARATION FOR DATA-LIMITED FISHERIES

A Visual Quick-Start Guide

Adapted from: Amoroso, R., Lomonico, S., Snouffer, B. & Gutierrez, N.L. 2024. Data preparation to inform assessment and management approaches in data-limited fisheries – A practical manual. FAO Fisheries and Aquaculture Technical Paper, No. 706. Rome, FAO. https://doi.org/10.4060/cd0366en

WHY THIS GUIDE?

In data-limited fisheries around the world, data are too often used at "face value" — fed directly into assessments or management frameworks without thorough exploration or review. This can lead to biased results and misdirected management. This guide distills the essential workflow for inventorying, cleaning, and exploring your fishery data before any formal analysis.

⚠ The Core Principle

Garbage in, garbage out. No assessment method — no matter how sophisticated — can compensate for poor-quality data inputs. The steps in this guide are not optional; they are the foundation of credible fisheries science.

THE THREE PILLARS OF FISHERY DATA

Data Type	Examples	Why It Matters
Biological Data	Lengths, weights, age, sex, maturity, fecundity	Provides life history parameters essential for stock status indicators, growth & maturity estimates, length-based assessments
Catch and Effort Data	Landings, total catch, trips, gear, soak time, fleet	Fundamental for understanding fishery trends over time; provide abundance indices (CPUE), exploitation trends, catch-based assessments
Traditional and Local Knowledge	Fisher interviews, expert knowledge, historical records	Validate data, fill gaps, understand fishery history, identify baselines. TEK/LEK may directly provide data on catch, effort, and biology; or be considered separately as valuable qualitative information.

THE 6-STEP WORKFLOW

Follow these six sequential steps to move from raw, scattered data to analysis-ready datasets. The process is iterative — expect to cycle back through earlier steps as you learn more about your data.

Importantly, these steps are focused only on inventorying and reviewing **existing** data streams. The steps to establish representative, unbiased sampling programs and methodologies are critically important, but not covered here.

1	2	3	4	5	6
Inventory Your Data	Build Your Timeline	Compile Life History	Set Up Your Database	Data Preparation	Exploratory Analysis

Each step is detailed on the following pages, with key questions, checklists, and practical tips to guide you through the process.

1 INVENTORY YOUR DATA

What: Create a comprehensive catalog of every data source that could inform the status of your fishery. This is a metadata exercise — you are documenting what exists, not analyzing it yet.

Why: Data are often scattered across institutions, formats, and time periods. Without a full inventory, valuable information may be overlooked, gaps may go unnoticed, and you cannot select appropriate assessment methods.

For each data source, document:

☐ Data type (catch, effort, length, abundance) ☐ Subcategory (e.g. reported catch vs. unreported) ☐ Source (landing site, market, onboard, survey) ☐ Fleet/sector and gear type	☐ Units of measurement ☐ Spatial coverage / area ☐ Temporal span (start year – end year) ☐ Data quality rating & notes ☐ Responsible institution ☐ Whether data are digitally available

? Key Questions to Ask

- Who else collects data on this fishery? (universities, NGOs, communities)
- Are there old paper records or logbooks that haven't been digitized?
- Are there other sources of information that might be useful (e.g. newspapers, historic accounts, local knowledge holders or practitioners)
- Do multiple sources cover the same variable for overlapping periods?
- What assessment methods may be feasible given the data I have?

Pro Tip: Consult Widely

Datasets are often decentralized. Engage fishers, community organizations, university researchers, and government staff early to develop a broad understanding of what information exists already. For example, unpublished theses, recreational fishing records, traditional knowledge hoders, and export statistics can all provide valuable information.

2 BUILD YOUR TIMELINE OF INFLUENTIAL EVENTS

What: Document events — environmental, regulatory, operational, or socioeconomic — that could affect the interpretation of your data over time.

Why: A sudden change in your data may look like a biological signal (e.g. recruitment failure), when it is actually caused by a regulation change, a new data protocol, or a hurricane. Without this timeline, you risk misinterpreting patterns.

Types of events to document:

Fishery & Management:	Environmental Events & Data Collection Changes:
☐ Fishery start date / catch peak ☐ Regulation changes (size limits, closures, quotas) ☐ Mesh size or gear changes ☐ Market or fleet changes	☐ El Niño events, hurricanes, habitat changes ☐ Changes in sampling protocols ☐ Shifts from paper to digital data collection ☐ Funding changes affecting monitoring

> **🔍 Example: The Mesh Size Story**
> Length composition data shows the sudden disappearance of small fish in 2016. A recruitment failure? No — cross-checking with the timeline reveals a new minimum mesh size regulation was implemented that year. The change in length composition is due to gear selectivity, not population dynamics.

3 COMPILE LIFE-HISTORY PARAMETERS

What: Build a repository of all available life-history parameter estimates for your target species, drawn from sources such as published literature, theses, global databases, FishBase, and local studies.

Why: Most data-limited assessments require life-history parameters as inputs (growth rates, natural mortality, maturity, length-weight relationships). Significant uncertainty can exist with life history parameters, especially when they are borrowed from other geographies or similar species. Although this uncertainty can dramatically impact results, analyses can be undertaken if uncertainty is acknowledged and accounted for.

Key parameters to compile:

Von Bertalanffy Growth: Linf, k, t0 — describes how individuals grow over their lifetime. **Natural Mortality (M):** Rarely measured directly; usually estimated from empirical relationships.	**Maturity (L50, L95):** Length at which 50% and 95% of individuals are sexually mature. **Length-Weight (a, b):** Relationship between body length and weight. Ensure units are standardized.

For each parameter, record:

The source reference, species sampled, sex (male/female/aggregated), life stage (mature/immature), geographic area, sampling year, sample size, parameter ranges (when available), and methods used. This metadata lets you evaluate which estimates are most reliable for your specific stock.

> **⚙ Sources of Variability**
> Parameter differences can arise from: (1) sampling variation, (2) true biological differences between sexes or regions, or (3) processing methodology. Visualizing the distribution of your data helps you spot outliers and plan sensitivity analyses.

4 SET UP & MANAGE YOUR DATABASE

What: Organize your raw data into a clean, consistent, rectangular or tidy format before any cleaning or analysis begins. Each type of data may require a separate, individual table.

Why: Poor database hygiene is one of the most common sources of error in fisheries analysis. Spreadsheet flexibility is a double-edged sword — it can make errors easy to introduce and hard to detect.

Essential database rules:

☐ One variable per column, one record per row ☐ First row = column headers (no spaces, no special characters) ☐ No formulas, graphs, or analyses in raw data files ☐ Store dates as separate Year / Month / Day columns ☐ Use "NA" for missing data; "0" only for true zeros	☐ Avoid color coding — add a column instead ☐ Use consistent species names (scientific + common) ☐ Use data validation / drop-down lists to prevent typos ☐ Create a data dictionary describing each column ☐ Prefer long format over wide format for analysis

Version Control Is Non-Negotiable

Never overwrite your raw data. If you are using Excel sheets, use “Save As” to create versioned copies (e.g. “Data_species_corrected_2024-03-15.csv”), or name each Version History in Google Sheets. Keep a log of every correction made between versions. Always maintain a pristine copy of the original raw data. Although the focus in this document is Excel, alternatively, using a coding software like RStudio can simplify the version saving process.

5 DATA PREPARATION (CLEANING & QA/QC)

Goal: Systematically review every column of your dataset to identify and correct errors, inconsistencies, and outliers. This is performed on raw data before any aggregation or summaries. Although the guidance in this document is intended for users who work in Excel, using a coding software like RStudio can simplify this process.

Phase A: Clean Text & Categorical Data

- List all unique values for each categorical column (species, gear, area, etc.).
- Look for: spelling errors, inconsistent capitalization, extra spaces, duplicate entries, and blank cells.
- Replace errors consistently and log every change.

Phase B: Clean Numerical Data

- Verify that all values are truly numeric (no stray characters).
- Check for impossible values (negative lengths, catches of implausibly large or small values).
- Validate date formats and check location coordinates.
- Ensure consistent units across the entire time series.

Phase C: Identify & Handle Outliers

Outliers fall into two categories that require different responses:

Error Outliers

Result from data entry mistakes, unit errors, or measurement problems.

Action: Correct if the true value can be determined; otherwise remove (replace value with NA) and document.

Interesting Outliers

Accurate data that reveal surprising or unusual real-world patterns.

Action: Keep and investigate. These may be some of your most valuable data points.

Outlier Detection Methods:

Method	How It Works	When to Use
Summary statistics	Min, max, mean, median, quartiles	Quick scan for impossible values
Z-score / Modified Z-score	Flag points > 3 standard deviations from mean	Works for symmetric data; modified version is more robust
Interquartile Range (IQR)	Flag points outside Q1−1.5×IQR to Q3+1.5×IQR	Robust; works well for skewed distributions
Visual inspection	Histograms, box plots, scatter plots	Most intuitive; essential for multivariate patterns

Critical Reminder: Univariate Is Not Enough

A fish length of 10 cm and a weight of 50 kg may each seem plausible alone, but together they reveal an impossible weight-at-length. Always follow univariate checks with multivariate exploration. Faceting plots by gear, area, year, and sex is essential for uncovering hidden patterns.

6 EXPLORATORY DATA ANALYSIS (EDA)

Goal: Learn from your cleaned data. Identify patterns, assess data coverage, cross-check multiple data sources, and test key assumptions — all before any formal statistical modelling or stock assessment.

Start with Data Coverage

Before interpreting trends, assess where and when you actually have data. Build coverage summaries across key grouping variables such as time, space, species, gear, and fleet. Identify gaps and determine whether the number of samples per group is sufficient to support analysis. The answer here depends on your reearch question and the variation in the data, but a basic starting point is 30-100 samples per group. For instance, if you want to understand whether the average capture length of *Lutjanus gibbus* varies across two sample locations, the number of samples required for will depend on how much variability is observed in the length data. Review the concepts of 'statistical power' and 'stratified random sampling' to learn more.

Your Visual Toolkit for Exploratory Data Analysis (EDA):

Plot Type	What It Shows	Type
Histogram / Density plot	Distribution of a single variable (e.g. length frequencies)	Univariate
Box plot / Violin plot	Spread, central tendency & outliers of a variable	Univariate
Scatter plot	Relationship between two variables (e.g. weight vs. length)	Bivariate
Faceted plots	Same plot split by categories (gear, area, sex, year)	Multivariate
Tile / Heat map	Number of samples across two dimensions (e.g. year × area)	Coverage
Time series plot	Trends over time (catch, effort, CPUE)	Temporal

EDA by Data Type:

Length Composition

- Plot length frequencies by species, year, area, gear, sex
- Compare to life-history parameter benchmarks (L50, Linf)
- Check if selectivity or sampling bias explains patterns
- Examine sex ratios and maturity proportions over time

Catch, Effort & CPUE

- Plot total catch and effort time series
- Cross-check official data with alternative sources
- Compare effort in different units (trips, days, hooks)
- Explore CPUE by fleet, gear, area, and season

? Key Questions to Ask

- Do trends in my data reflect real biological/ecological changes, or are they artifacts of how data were collected?
- Is my data coverage representative of the fishery in time and space?
- Do different data sources tell a consistent story? If not, why?
- What assumptions about my data need to hold for my planned assessment or indicators to be valid?

REMEMBER: THIS IS ITERATIVE. Data exploration is not a one-pass exercise. Exploratory data analysis will uncover new issues that send you back to data preparation. Data preparation may reveal gaps that require updating your inventory. Each cycle deepens your understanding and strengthens the foundation for any assessment or management decision that follows.