WIOMSA AI Workshop: Background

AJ Smit

Table of contents

1 Main Objectives

- **Part One – Data Collection:** Identify a reliable source of textual data in marine science (WIO-focused) and obtain the content (plain text or PDFs). Guide participants to process the text (reading, summarising, or extracting specific facts) and build a dataset or report from it.
- **Part Two – Data Processing:** Here the emphasis will be on AI-assisted data analysis through the use of their scripting language of choice (R or Python).

(*Provide optional paths or extensions (e.g., adding more sources or advanced AI tools) if students advance quickly, and simplifications or support if needed.*)

2 Part One – Data Collection (Textual Data Sources)

This is the confidence-building phase, so we will want to progress from simple to more complex tasks.

Start with a well-known, trusted publication that is rich in relevant information. This ensures content quality and reduces the chances of technical issues during processing (thus offering a "fail-safe" approach). Below are sources I'm exploring… they focus on the WIO region, and some of which could be used for Part One. Each is authoritative, freely available, and well-suited for text analysis:

- **UNEP's Regional State of the Coast Report (WIO):** The *Regional State of the Coast Report: Western Indian Ocean (2015)* comprehensively covers the WIO region. It provides a broad synthesis of topics – from fisheries and coastal resources to climate change, biodiversity, and policy scenarios. It contains multiple chapters of text and is ideal for extracting many types of information. Being a United Nations Environment Programme publication, it's well-known and credible, and it specifically addresses WIO countries. Workshop participants can parse this report

to learn about regional fishery status, coral reefs, mangroves, socio-economic factors, etc., all in one document.

- **WIO Marine Protected Areas Outlook (2021):** This is a joint Nairobi Convention–WIOMSA report that details the status of marine protected areas in each WIO country. It's a region-specific source for conservation and policy data. For example, the Outlook documents that the WIO region has 143 MPAs covering about 555,000 km² (~7% of the region's EEZ), and discusses progress toward global biodiversity targets. If the participants are interested in conservation assessments, this report provides ready-to-use textual data on ecosystem protection efforts, which can be analysed per country or region-wide.

- **Coral Reef Status Report for the WIO (2017):** The report is published by the Global Coral Reef Monitoring Network (GCRMN) and ICRI. It is another high-quality text source focussing on ecosystem health. It contains detailed findings on coral reefs in the WIO after major bleaching events. For instance, it reports that average live coral cover in the WIO declined by about 25% after the 1998 global bleaching event; this fundamentally alters reef fish communities. If our focus leans toward climate change impacts on ecosystems, participants can extract trends, data points, and regional comparisons from this document.

I'd suggest starting with one primary source – ideally the UNEP State of the Coast Report (WIO) for its breadth – to keep things manageable. This report is well-suited to Part One because it's a single coherent document covering multiple relevant themes in the WIO. Participants can be divided into groups to approach different chapters or topics from the report. As confidence grows, we can introduce additional sources (for example, an MPA Outlook chapter or a specific FAO report section) to enrich the dataset. All suggested sources are from well-known organisations (UNEP, WIOMSA, IUCN, FAO) and ensures the information is trusted and of high quality. Lastly, these publications are publicly available.

2.1 Stages of Processing
The following steps are broadly covered in the above framework:

2.1.1 Stage 1 — Demonstration of AI as a "Text Assistant"
- Task: Upload a section of the UNEP State of the Coast Report (WIO) into ChatGPT (or a similar model) and ask it to:
  ‣ Summarise key findings.
  ‣ Extract species names, locations, or trends.
  ‣ Translate a passage into another language spoken in the region (e.g., Swahili, French, Portuguese).
- Confidence-building element: Participants quickly see how AI makes a dense scientific report more accessible, without needing any coding.

2.1.2 Stage 2 — Prompt Engineering for Marine Science
- Inspired by the workshop PDF:
  ‣ Show how a basic prompt ("Summarise this fisheries report") produces a generic response.

‣ Refine prompts iteratively ("Summarise in bullet points, focusing on coral reef health, and identify all species mentioned").

‣ Emphasise control and precision — demonstrating that participants can direct the AI, not just passively consume output.

2.1.3 Stage 3 — Structured Outputs for Research Use

• Task: Give AI both a text excerpt and a desired output schema (e.g., "Produce a table with columns: Species, Distribution, Dietary items, Threats").

• AI attempts to parse into structured data. Participants review outputs and correct inconsistencies.

• This mirrors the "merge prompt with example output" idea I have previously used in another workshop — teaching participants to combine task + format to get useful, reproducible results.

2.1.4 Stage 4 — Group Exercise: "Building a Mini-Database"

• Each group gets a short passage (from UNEP, FAO, or WIOMSA text).

• Using AI, they extract structured facts (species, regions, stressors, MPA stats, etc.).

• Groups then combine outputs into a single group-wide CSV file.

• Novices can copy-paste and run simple AI queries, while advanced participants can start thinking about validation and cross-checking with the source text.

3 Part Two – Processing and Analysis (Building the Dataset)

Once the textual data is collected in Part One, Part Two will focus on processing that text and extracting the information we need:

• **Text Extraction:** Most GPTs can handle PDFs directly (pre-processing tools might help with more difficult resources, but I don't think it's an issue here). Given the large volume of pages in the recommended sources, we might split the text by sections or chapters to distribute the workload. For example, Chapter 5 of the State of Coast Report might deal with fisheries, Chapter 7 with coastal habitats, etc., which different teams can handle.

• **Defining Data Fields:** Decide what information will go into the final CSV (or chosen output). This depends on the project's focus:

‣ If we follow the fish ecology example, relevant fields might include Species Name, Distribution (geographic range in WIO), Feeding/Diet, Growth Parameters, Conservation Status, etc.

‣ If focusing on climate change impacts, fields might be Location, Observed Impact (e.g., coral bleaching extent), Year, Adaptive measures, etc.

‣ For conservation/MPAs, fields could be Country, Number of MPAs, Total Area Protected, Percentage of EEZ protected, Notable Species Protected, etc.

We will keep this schema flexible – participants can propose additional fields or simplify based on what the text actually contains. The key is to encourage critical thinking about how to structure unstructured information.

• **Manual vs. Automated Extraction:** To ensure a "fail-safe" progression, start with a manual or semi-structured approach: have students read through the text and identify key pieces of information for a few entries. For example, they might manually tabulate the data for one country or

species as a pilot. This helps them understand context and verify that the information is correctly interpreted (important when dealing with nuanced scientific text).

If the group of participants is more advanced or once they grasp the basics, we can introduce automation:

- Use keyword searches or regular expressions on the text to find lines with specific data (e.g., "growth rate", "distribution:").
- Employ simple NLP tools or AI assistants. Given the interest in AI (as per our survey), one could use a language model to extract info. For instance, students might prompt an AI (GPT or similar, if available) with a passage: "Extract the species name and its diet from this paragraph." However, emphasise cross-checking any AI-extracted data against the source text for accuracy.
- *Advanced, optional:* If the resources allow, using a tool like Knime or Orange (visual workflow tools) or a Python script to systematically pull out data for all entries could be a nice extension task. This can be done once students have confidence that the approach works on a small scale.
- **Building the CSV:** As information is extracted, compile it into a CSV file (or Google Sheets/Excel for easy viewing). Ensure everyone uses a consistent format and agreed-upon field names. This is a good exercise in data standardisation. For example, if one group writes "Habitat: coral reefs" and another writes "Primary Habitat – Coral reef", the we should reconcile these to a single schema (AI may be used for this). Maintaining a shared template or example row can help.
- **Quality Control:** Before finalising the CSV, do a round of verification:
  - ‣ Spot-check a few entries against the source documents to confirm accuracy.
  - ‣ Look for any obvious outliers or inconsistencies in the data (e.g., a fish growth rate that seems off by an order of magnitude, or a country listed twice with slightly different names).
  - ‣ This QA step can be done by peer review, so groups swap a portion of their extracted data and validate each other's entries against the text.

Throughout Part Two, adapt to the participants' pace. If they struggle with extraction, we can slow down and perhaps reduce the number of fields or entries. Conversely, if they finish quickly, an extension could be to incorporate additional sources from Part One or to do some basic analysis on the CSV (like generating summary statistics or charts from the data – e.g., average % of protected area across countries, or a bar chart of number of fish species by region, depending on the data collected).

3.1 Stages of Analysis
(To be developed.)

4 Adapting to the Group's Progress
We need to maintain flexibility. Here are ways we can adjust on the fly:

- **Scope of Data:** Begin with a narrow scope (e.g., one report, or one theme like "coral reefs") to ensure everyone can participate meaningfully. If all goes well, broaden the scope – add another document or another theme (for instance, after covering coral reefs from the State of Coast report, we might introduce a fisheries chapter or the MPA Outlook data for comparison). This modular approach means the project won't overwhelm the group at the start, but there's room to grow.

- **Depth of Analysis:** The initial goal might simply be to summarise each section of text or list key facts. For example, students could start by writing a short summary of what the State of Coast Report says about fisheries or climate change. This ensures comprehension of the material. Only afterward would we move to the more technical task of structuring that information into a CSV. If the group is not ready to create a structured dataset, staying at the summary level is still a valuable outcome (they learn from the content and practice synthesising information). On the other hand, if participants are comfortable, we can push deeper into data extraction and maybe even light analysis of the compiled data.
- **Incorporating AI Tools:** Decide how much to integrate AI assistance based on the group's skill. A looser approach could be letting students experiment with an AI chatbot on a small scale (e.g., "Ask the AI to extract one specific fact and see if it matches the text"). If it works reliably, they can use it more; if it proves confusing or inaccurate, we rely more on manual methods. The project can succeed with or without heavy AI usage, so we can adapt this up or down. The emphasis should remain on understanding the content and the process, rather than just getting an answer from a model.
- **Feedback and Iteration:** After Part One (data gathering) and initial extraction trials, we will discuss challenges faced. Maybe the text was too dense in places, or some data points were hard to find. We can then adjust the plan – for instance, if the State of Coast Report is too broad, we might shift to a more specific source (like focusing only on the Coral Reef Status Report for clarity), or *vice versa*. The loose structure means this check-in and shift is built into the schedule. It's okay if we don't stick rigidly to an original outline as long as learning objectives are met.
- **Optional Extensions:** If their progress fast, consider adding an extension task such as:
  - *Visualising* the extracted data (e.g., creating a simple map of WIO countries with MPA coverage percentages, or a graph of coral cover decline over time if we got that data).
  - *Writing a short insight report…* each group can draft a paragraph interpreting the data they compiled (e.g., "Country X has the largest protected area coverage at Y%, which might be due to Z")… this can be taught together with AI ethics concerns. This blends data processing with critical thinking and ensures they don't see the CSV as just numbers but as something with real-world meaning.
- *Comparing multiple sources:* If we have time to use two sources (say the UNEP report and a FAO report), an interesting exercise is to see if they tell consistent stories. Do the numbers or statements in one source align with the other? This can lead to discussions on data reliability and cross-verification in research.

Throughout the project, maintain liberal and free communication with the group. This is where the floor assistants will come in and play a key role. If a particular step is too time-consuming or technically difficult, we can simplify it (I'll have to adapt on the fly, somehow). The goal is for the participants to learn about marine science in the WIO and gain experience in handling real-world textual data, not to get stuck on a rigid procedure.

Bibliography