DIGITIZATION OF SUSTAINABILITY DATA

Technical approaches ("best practices") relating to data modelling for collecting and managing sustainability information

TPFIII & IMP Workstream B

March 31, 2022

Table of Contents

Abstract: 2

Problem Statement: A Sea of Challenges 2

Objectives: 3

Hypothesis: 3

- **1. Data Supply Chain Analysis:** 3
- **2. Controlled Vocabulary:** 3
- **3. Metric Harmonization:** 3
- **4. Data validation in supply chain:** 3

Procedure (Methods and techniques): 4

- **1. Supply Chain Analysis:** 4
- **2. Controlled Vocabulary:** 4
- **3. Metric Harmonization:** 5
- **4. Data validation in supply chain:** 7

Observations: 8

Results: 8

Conclusion: 9

Appendix Material: 10

- **Seed Document Links:** 10
- **Website Links:** 10
- **Federal Register Link:** 10

Abstract:

The current sustainability information value chain is fraught with gaps and issues that limit decision-usefulness. The information flow system lacks FAIR data, that is data which embeds findability, accessibility, interoperability, and reusability. More recently, regulation is now embedding a requirement to further assure the data veracity. It all adds up to a global sustainability information ecosystem that is not fit for purpose in a digital age.

This workstream set out to demonstrate how technology, Knowledge Graphs and machine learning, can create a game changing trusted information flow system. Our prototype is designed to demonstrate a model for efficiently and effectively defining & understanding data requirements, and controlled vocabularies. The ultimate intent of which is to provide a consistent digital way to describe data in open data models which ensure inclusion for all stakeholders within the local and global economies.

Problem Statement: A Sea of Challenges

The ocean faces significant risks including overfishing, biodiversity loss, pollution from terrestrial land use change, ocean acidification, habitat destruction, spread of alien invasive species - to name just those associated with natural capital. Significant social and human capital challenges are associated with this degradation of natural ecosystems, including health, wellbeing, and livelihood loss for stakeholders from coastal communities to countries and large corporations. The ocean is vast, deep and there is a great deal we still don't know! Some of the issues include overfishing, by-catch, IUU - Illegal, Unregulated and Unreported fishing, Seabed damage from bottom trawling and plastic and fish net pollution.

The Achilles' Heel of the industry is the lack of consistent and comparable data which disrupts the flow of information from ocean (origination) to plate (consumption), impacting all decision-making along the stakeholder value-chain (figure 1).

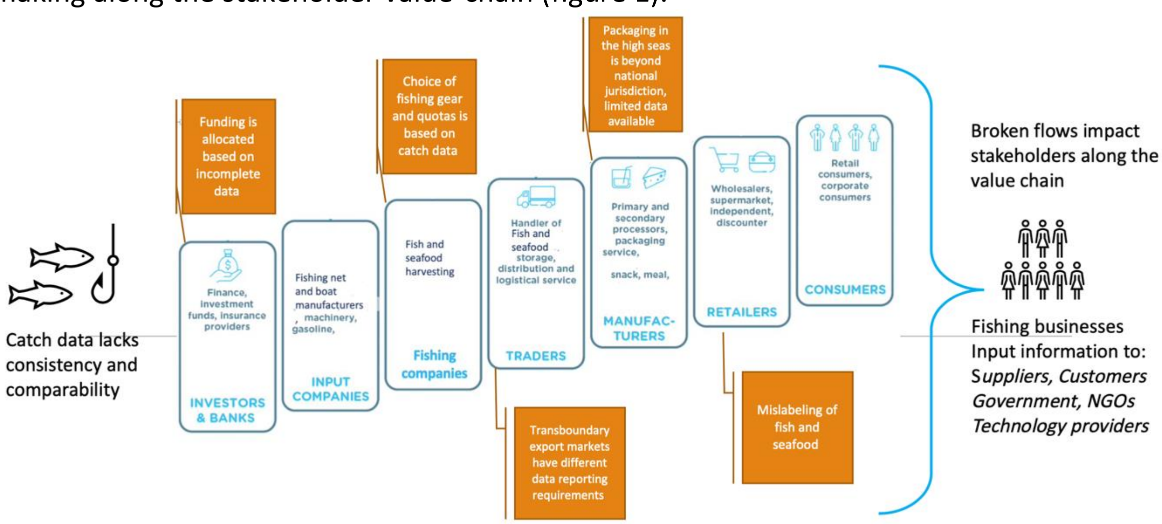

Figure 1: Simplified Information Value Chain – Fisheries

We need digitized information flow solutions that will enable meaningful decision making. The hypothesis we seek to test is that technology, specifically Knowledge Graphs and Machine Learning (herein referred to an Augmented Intelligence Platform, AIP), can accelerate the journey towards a fit for purpose digital information flow solution.

Objectives:

1. Identify data quality issues and challenges within the fisheries data supply chain, determine possible (probable) root causes, and suggest remediation steps using an augmented intelligence platform
2. Build a controlled vocabulary that systematically links concepts, components, metrics, a knowledge graph, and augmented intelligence on fisheries supply chain
3. Harmonize metrics across different sustainability standards using augmented intelligence
4. Validate all recorded transactions within the fisheries data supply chain

Hypothesis:

1. Data Supply Chain Analysis:

The Augmented Intelligence Platform (AIP) can Identify key data quality issues within the fisheries data supply chain and suggest remediation actions.

- AIP can ingest, analyze, and process fisheries supply chain documents, and identify the key issues/challenges at each phase within the supply chain.
- AIP can identify and recommend remediation actions and identify appropriate metrics to efficiently track the issues in the future.

2. Controlled Vocabulary:

AIP, in conjunction with a subject matter expert's guidance, can construct controlled vocabularies for the target domain. The resultant vocabulary can be generated, and visualized, in a user-defined manner (structure).

3. Metric Harmonization:

AIP can cover aspects of sustainability and multiple sustainability standards, coupled with subject matter expert's knowledge, to achieve high quality harmonization of metrics within hours (rather than weeks or months for a subject matter expert unsupported by an AIP).

4. Data validation in supply chain:

AIP can analyze the data at each phase in the supply chain and verify data authenticity. It can flag any anomalies and transactions that deviate from standards or expectations.

Procedure (Methods and techniques):

1. Supply Chain Analysis: Every supply chain experiences challenges to achieve a successful and orderly operation. Even within an efficient supply chain, there may be aspects misaligned with sustainability standards. For instance, identifying the key issues within a fisheries supply chain and resolving those through appropriate remediation activities, will be a significant step towards the entire supply chain becoming sustainable.

The fisheries supply chain is divided into four functional phases:

1. Point of catch
2. Point of processing
3. Trade and transportation
4. Purchase and consumption

Each phase includes a list of concepts from the domain glossary, which the AIP has learnt (ingested). Each concept includes related metrics across all ingested standards and its data. Additionally, the machine exposes any identified data issue within the fisheries supply chain phase. In addition to the issues are the associated metrics, identified by the AIP, which aid issue tracking.

2. Controlled Vocabulary: Taxonomy creation begins with focused domain documents supplied by a subject matter expert (SME). In the fisheries prototype, we built the AIP starting with four seed documents, eight website data (links) and select federal register documents. The content extracted from these documents, repositories and websites provided the aggregate foundation data learning for the AIP. The machine ingested these documents to learn about the defined scope fisheries domain. A more "sophisticated" understanding by the AIP was achieved by expanding the initial document set to allow the AIP to develop greater granularity. Once the foundation (document seeding) is completed, the AIP self-directs the additional download of relevant documents from public sources and builds a coherent corpus (real life text) through its built-in filtering mechanism. In this case, the AIP generated approximately 2000 documents based on the seed data ingested. Once built, the AIP generated the domain concepts, components, and their connections which were stored in a knowledge graph. For this prototype run, the AIP identified 140,000 concepts and 710,000 unique connections.

A knowledge graph illustration, below in Figure 2, of the "Red Snapper" visualizes both the concepts and the connections for ease of understanding.

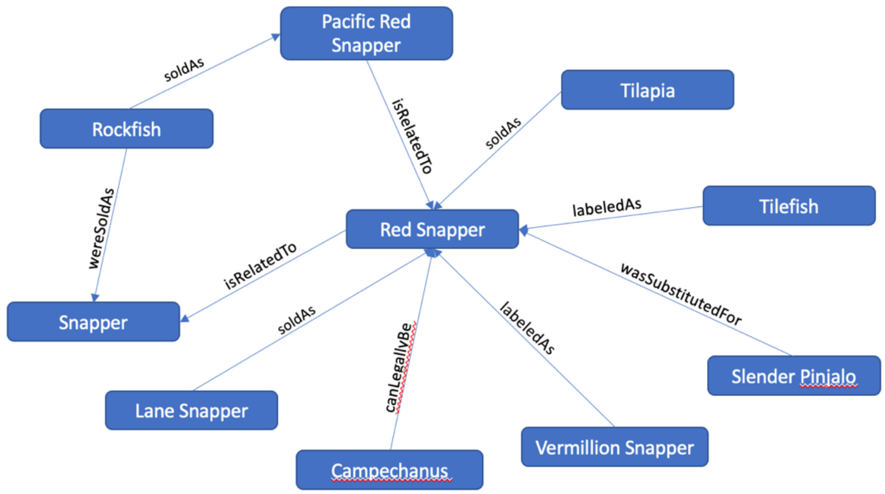

Figure 2: Concept Connections

Important concepts and connections can further be shortlisted based on their scores and connection strengths, respectively. The SME makes the final call on "dialing-in" the threshold for scores and formally approving the controlled vocabulary. We added an additional 40 controlled set of documents to extract components of multiple dimensions, and this was appended easily with the controlled vocabulary, when required. Knowledge graphs are powerful because they contain a logical and documented understanding of each concept (a detailed signature) within the domain.

3. Metric Harmonization: Each enterprise follows required standards, based on their geography and industry type, for instance, amongst other things. Multinationals are required to adhere to multiple standards given their multijurisdictional reach. Data elements extracted from structured and unstructured sources are associated with metrics relating to individual standards. To extrapolate this exercise for multiple standards manually is, as one can imagine, time and resource intensive, and prone to human error. In contrast, AIP associated data elements with a given, or across multiple standards, harmonize across all available standards "instantly" with a high degree of accuracy.

The AIP generated a complete and custom signature of several dimensions for each metric, a complex web of connections to the human eye. Metrics across standards were mapped based on the similarity of their signatures. Figure 3 below illustrates an example of the multi-dimensional signature of a metric and its connections to metrics across multiple standards.

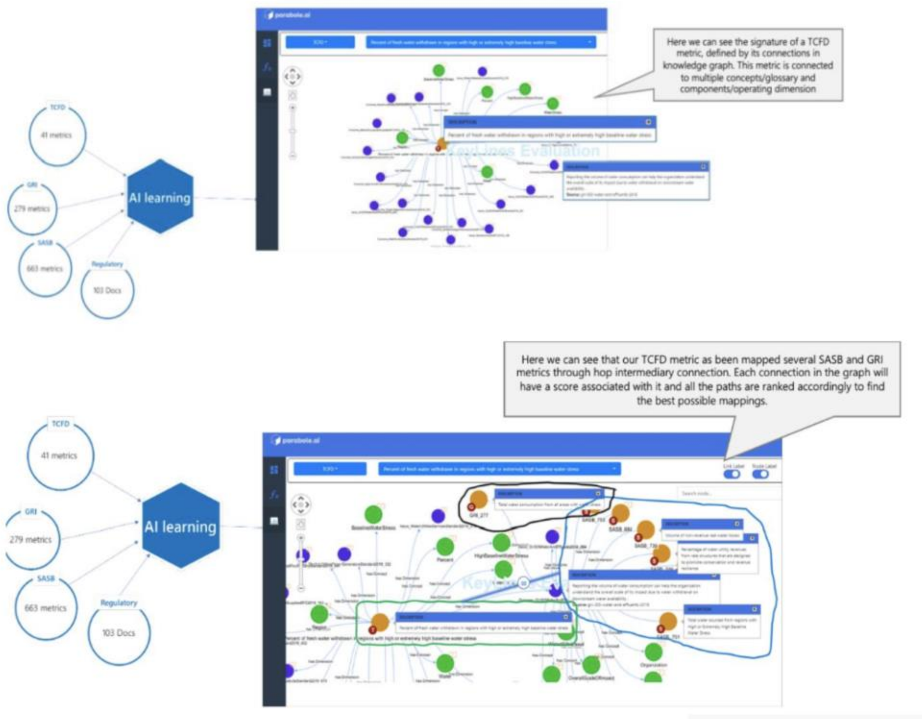

Figure 3: Metrics Harmonization

During the prototype, the AIP machine ingested 3 different standards (SASB, GRI and TCFD). The controlled set of documents used generated accurate and relevant components (connections) of multiple dimensions, along with the most relevant controlled vocabulary. These components and the controlled vocabulary were essential for mapping metrics across standards effectively. The SME played a key role in validating the mappings, which reinforced the AIP model's effectiveness.

To ensure ease of processing, the user may upload any new standard information to be harmonized with all AIP previously ingested standards within a matter of mere minutes.

Figure 4 below illustrates the user interface (UI) for adding a new standard into the AIP environment.

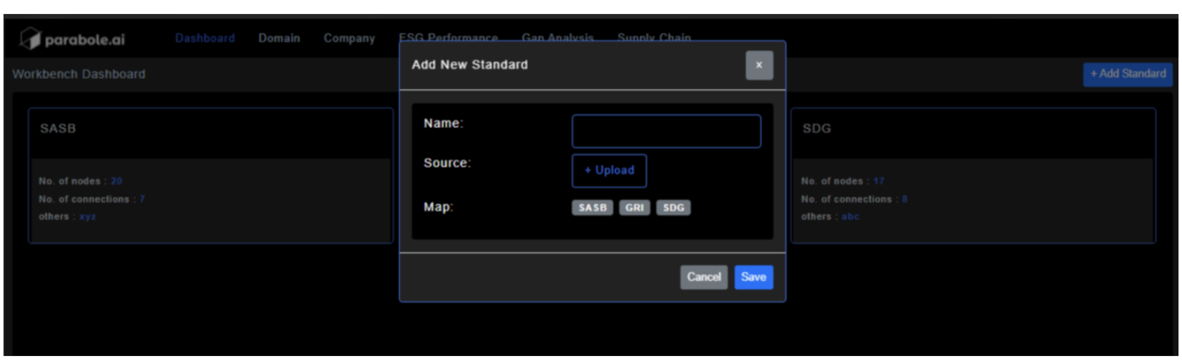

Figure 4: UI for adding new standards

Once the standards data is ingested, users can upload company specific documents to visualize the reported data in accordance with required standards. Figure 5 below illustrates the user interface for adding company specific documents into the AIP environment.

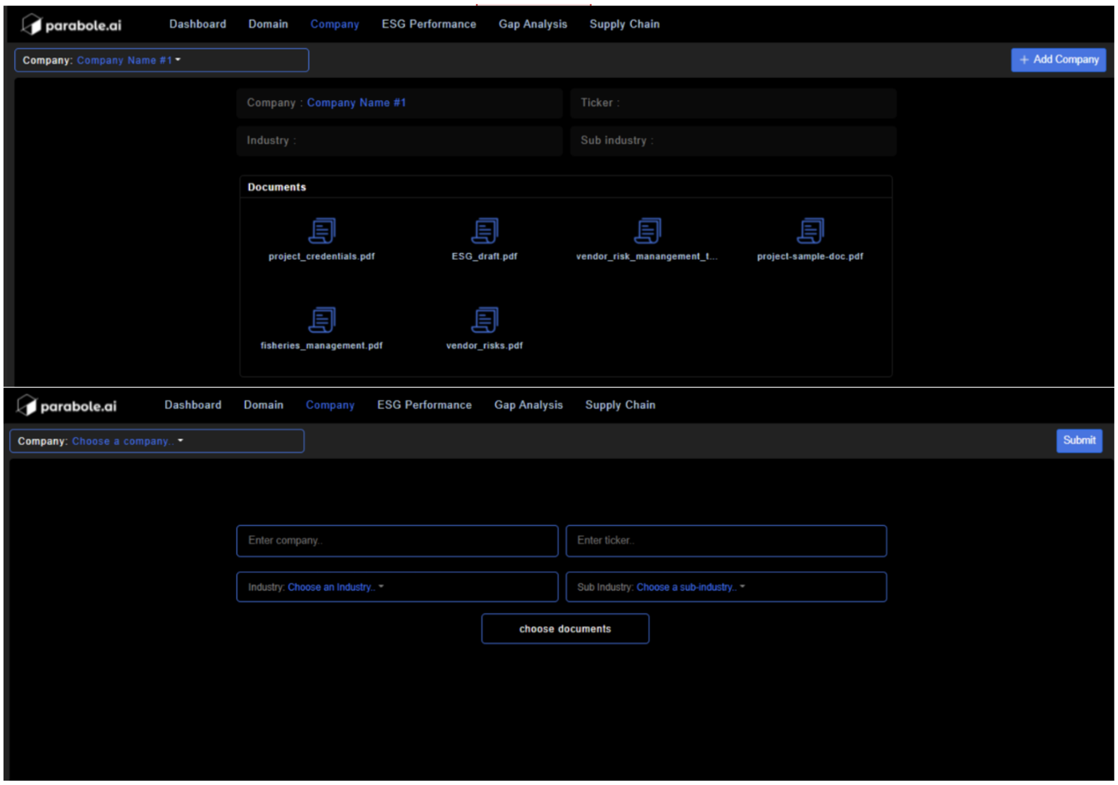

Figure 5: UI for adding company specific data

4. Data validation in supply chain: It is important to analyze whether the data received at each of the phases in the supply chain is authentic. For example, when a particular species of fish is recorded at the "point of catch" phase with a defined weight and in a certain location, the AIP analyzes the transaction and validates it. Any abnormal data points (including those originating from data quality issues) will be flagged for action.

Within the prototype, such data anomalies (see Figure 6) were identified when the user reported a catch of 1,000 lbs. of herring in the Irrawaddy River. The machine flagged the transaction, based on its intelligent knowledge.

To explain the first flag: herring is a marine species whereas the Irrawaddy is a freshwater body. The AIP deduced, based on geographic and species information that the most likely species of the catch is Hilsa - the dominant species in that geographic location.

A second flag was created when the catch was transaction includes the catch of the Hilsa species during the month of October. This transaction was flagged because the AIP identified a

Hilsa catch prohibition during October. Why? October and November are the breeding season for the Hilsa species and therefore fishing is prohibited.

The validation of data goes beyond the point of catch phase and carries through the entire fisheries value supply chain ensuring decision useful information which promotes a more sustainable fisheries supply chain.

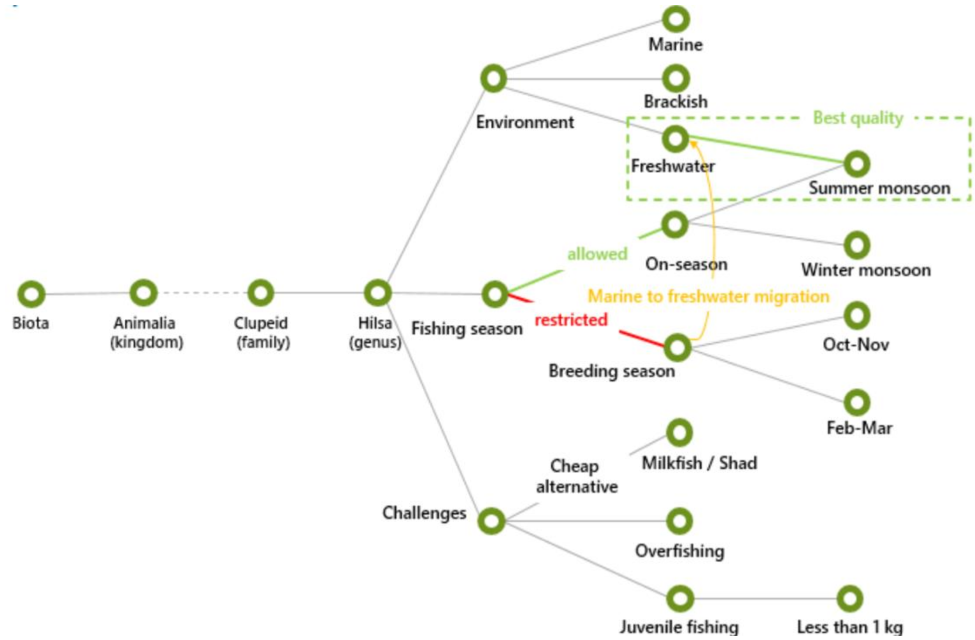

Figure 6 : Data anomalies (Causal inferences)

Observations:

The AIP depends on the knowledge of SMEs to learn "correctly", and taxonomy connections requires multiple iterations to achieve a high-quality output. Secondly, metric harmonization requires high-quality documents covering all the aspects of sustainability. The quality of harmonization is significantly dependent on the quality of input documents. Finally, genuine supply chain documents need to be ingested to verify the AIP's accuracy in identifying key issues/challenges, and in its accuracy to systematically validate data.

Results:

1. Identifying issues and challenge in fisheries supply chain is completely unsupervised. AIP's suggestions of remediation steps were validated, and the SME's feedback improves the overall inferencing accuracy.
2. Within the limited scope of the prototype experiment, the AIP process of controlled vocabulary generation demonstrated a high-quality output. With an SME's limited curation, it achieved a 90% accuracy after only a couple of iterations. Testing involved SME's validation of

each concept in the vocabulary and its first level connections. Precision and recall are calculated based on the SME's review of the final vocabulary.
3. Metric harmonization was rapid with a high degree of accuracy, needing minimal supervision to achieve peak performance.
4. Data validation was completely unsupervised. The basic challenge in the prototype was the provision of comprehensive data across all four phases, stakeholders and entities involved within the supply chain. The quality of output correlates to the completeness of the data provided.

Conclusion:

1. The best approach to digitize sustainability information flows and instil "Trust" in decisions and actions requires "Human Centric AI" rather than exclusively relying on AI that is prone to model biases, also known as Model Toxicity.
2. The AIP prototype identified issues/challenges within the fisheries supply chain and flagged transactions with data anomalies, which otherwise would be a complex labour-intensive task. In addition, the prototype further suggested remediation actions paving the way for a more sustainable supply chain (through better decision-making).
3. Augmented by human intelligence, the prototype acted as an accelerator in building and maintaining a controlled vocabulary and in harmonizing multiple standards (as demonstrated).

Effectively, the prototype demonstrates the efficacy of leveraging machine intelligence as an accelerator, scalable method to interpret, identify patterns and harmonize sustainability information and Taxonomies. Machine Intelligence connects a multitude of information sources to highlight "anomalies" or "out of context", meaning contextually conflicting, Insights. The additional validation by human subject matter experts renders "Supervised AI" more pragmatic in accuracy and effective in informing decision making.

Appendix Material:

Seed Document Links:
https://medblueconomyplatform.org/wp-content/uploads/2021/03/2021_turning-the-tide-guidance_un-environment-programme.pdf
https://www.oceanpanel.org/ocean-action/files/full-report-ocean-solutions-eng.pdf
https://oceanpanel.org/sites/default/files/2019-11/19_HLP_BP1%20Paper.pdf
https://reliefweb.int/sites/reliefweb.int/files/resources/The%20State%20of%20World%20Fisheries%20and%20Aquaculture%202020.%20In%20brief.pdf

Website Links:
https://www.fpilab.org/fpi-home/
https://abalobi.org/
https://www.msc.org/standards-and-certification/fisheries-standard
https://friendofthesea.org/sustainable-standards-and-certifications/sustainable-fisheries-and-fleets/
https://www.gov.uk/guidance/fishing-certification-and-training
https://www.audubongulf.org/certification/about-g-u-l-f-certification/
https://www.worldbenchmarkingalliance.org/publication/seafood-stewardship-index/
https://www.fao.org/documents/card/en/c/ca9229en/

Federal Register Link:
https://www.federalregister.gov

Contributors to this report

Tanuj Agarwal, Bon Conseil
Tom Baumann, Climate Check
Eric Bigelsen, EDM Council
John A. Bottega, EDM Council
Nilesh Dadi, Microsoft Corp
Kathy Dubansky, Workiva, Inc
Kevin Dutt, Princeton University
Pedro Faria, CDP
Isabel Hoffman, Capitals Coalition
Urmish Mehta, Bon Conseil
Jaume Mora, Engine B
Javier Mora, Adviseers
Michal Piechocki, Business Reporting – Advisory Group
Charan Pothireddi, Parabole
Shaun Rolls, EDM Council
Rajib Saha, Parabole
Harsh Sharma, Microsoft Corp
Adrian Van der Merwe, Sonum International
Liv Watson, Adviseers, SAS
David Wray, DFCG