Odini The Ocean Data Integration Initiative

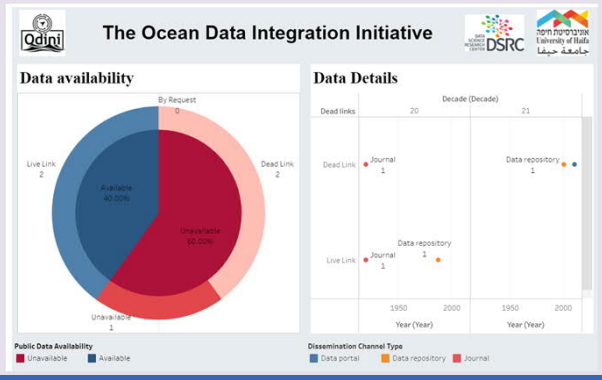
Taking Stock of Decades of Ocean Data

Tomer Sagi[1] & Yoav Lehahn[2]
[1]Department of Information Systems [2]Department of Marine Geosciences, Charney School of Marine Sciences, University of Haifa, Haifa, Israel.

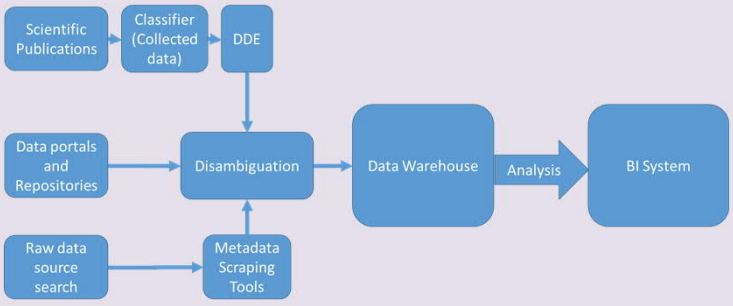
Background and motivation:

Oceanographic research relies heavily on the collection, analysis, and interpretation of data. Many papers in the various domains comprising oceanography begin with a statement such as "the amount of data available is steadily increasing". However, to the best of our knowledge, no one has taken a longitudinal approach to reviewing the availability, coverage, and amount of research data collected and published. We have began a systematic effort to collect and analyze mentions and records of data collection throughout the history of oceanographic science to quantitatively questions such: How much of the data collected is still available for analysis? **In which disciplines is data available over all regions of the ocean and in which are there shortages of available data?**

We further intend to make public an data analysis tool, allowing faceted exploration of questions such of these over the results of our work.



Business intelligence (BI) has been used for decades to organize and present information regarding the business environment of an enterprise and its internal performance measures. BI Systems are comprised of a data warehouse (DWH) that collects faceted information over measures (e.g., sales, profit, service calls) and dimensions (e.g., years, product-types, customer-types) and an analysis portal that allows managers and analysts to derive insights from this information. We propose a **Research Intelligence** system comprised of a similar DWH and analysis portal.



The diagram on the left presents our pipeline that is comprised of two major extraction, transformation, and loading (ETL) processes.

Data portal ETL

1. Extract metadata from data portals such as EMODNet, DataOne, and PlanetMicrobe, data repositories such as BCO-DMO, PANGAEA, and PODAAC, and raw data sources such as ARGO and BATS.
2. Transform extracted metadata into a common schema for loading into the DWH.
3. Disambiguate duplicate records (e.g., https://doi.pangaea.de/10.1594/PANGAEA.79008 and https://doi.pangaea.de/10.1594/PANGAEA.79051), overlapping datasets, and records fully contained in other records.

Evaluation procedure – scientific papers:

1. Use NLP-based topic-classification techniques to identify oceanographic papers with data mentions.
2. Utilize data description extraction (DDE) [1] to extract the data mentions and their properties.
3. Disambiguate against our collected data warehouse.

Preliminary Schema

The figure on the right presents a snowflake diagram of our current data warehouse design. As additional sources are incorporated. We intend to evolve the schema as needed. The center table represents the main data table and utilizes 3 measures: number of data points, number of datasets and number of papers. These measures can be grouped and filtered using the dimension hierarchies surrounding the fact table such as by month➔year➔decade or by depth (Z)➔ocean layer.

The list of data sources we are currently evaluating for this work can be found on our website at https://odini.net/discover