Reusability of data with complex semantic structure

Michal Kucera[1], Robert Huber[1,2]
[1]MARUM - Center for Marine Environmental Sciences, University of Bremen
[2]PANGAEA - Data Publisher for Earth and Environmental Science

Date of submission: 30.7.2020

Abstract

Data on the occurrence and abundance of fossils provide invaluable insights into past climates and past ecosystem response to perturbations. Since such data are generated with substantial operator input (taxonomic identification) and use complex vocabularies (names of taxa, changing in time and inconsistent among operators), their reusability is severely limited, hindering global syntheses as a basis for global assessment of biotic response to climate change. Here we propose to facilitate reusability of new and old fossil occurrence and abundance data by developing a community-based workflow. Using data on Quaternary planktonic foraminifera as a model, we will combine community-driven development of semantic standards with technical implementation making use of NFDI core service platforms. We intend to involve selected researchers, learned societies, data scientists and data providers with the overarching aim to develop and demonstrate an approach to enhancing reusability of data with complex semantic structure that could be transferred on different types of long-tail data within the broad remit of NFDI4Earth. The approach considers upfront the full data life cycle, allowing integration of legacy data with workflows for new data submissions, and provides an intersection to NFDI4BioDiversity, thus fostering synergy within the broader NFDI community.

I. *Introduction*

The occurrence of living organisms reflects their adaptations to environmental conditions and their fossils preserved in the geological record can therefore be used as a source of information on past environments and climate and on the reaction of past ecosystems to perturbations (Yasuhara et al., 2017). Large amounts of data on the occurrence and abundance of fossils have been collected over decades, if not centuries, but because of the complex semantics, reflecting the intricacies of biological nomenclature and its inconsistent application by users, the resulting treasure trove of paleoecological data is hard to mine for global patterns. This situation can be exemplified by data on the composition of sedimentary assemblages of planktonic foraminifera, prolific marine microplankton with excellent and richly studied fossil record, for which the proponents have the necessary scientific expertise and professional network. These data stand for only a small section of the diversity of fossil organisms, but already for this group, the amount of available data exceeds the threshold for manual curation even for data originating from a single time slice (Siccha and Kucera, 2017). Using a semi-automated pipeline to process data lodged in public repositories, Siccha and Kucera (2017) generated a globally consistent resource, which have been used to show how modern plankton ecosystems departed from their pre-industrial state (Jonkers et al., 2019) and how tropical marine diversity may decline under future warming scenarios (Yasuhara et al., 2020). These examples are just scratching the surface of the potential

of the existing data: analysing global patterns of biotic response to climate change, such as rates of assemblage turnover and range expansion, changes in the structure of trophic webs, functional consequences of declining biodiversity or the origin of modern communities all require access to syntheses across many taxa and time scales.

A large portion of the fossil occurrence and abundance data is publicly available in repositories, but because of their complex semantic structure, the data are not reusable without manual curation and expert knowledge. The same problem applies to many other types of earth science data and as a result, preparing data and ingesting them into analysis platforms (e.g., for statistics and models) consumes a large share of the required resources to analyze the data by researchers. Even within highly organised data archives, small structural differences and semantic incompatibilities compromise the reusability of the data. This is in particular the case when highly specialized vocabularies (e.g., taxon names) are used, which frequently change or exist in variants expressing uncertain identifications of the scientific subject (e.g., open nomenclature) and where the application of identifiers is inconsistent among the users.

As a result, taking the example of marine microfossil occurrence and abundance data, the scientific community proceeds by periodically generating data synthesis products. This practise has many undesired consequences, beginning with substantial efforts needed to dereplicate the various syntheses (Siccha and Kucera, 2017), over lack of standardisation in the generation of the syntheses leading to loss of metadata and finally ending with the necessity to periodically update the syntheses, which become soon outdated because of continuous generation of new data. The existence of the fossil data in the form of many isolated datasets with disparate ontologies (long-tail data) hinders their usage across user domains as well as their interconnection with other data generated from the same geological samples, which is a common problem in Earth sciences (Jonkers et al., 2020).

The lack of community consensus on vocabularies and the lack of a suitable workflow associated with archiving of Earth science data with complex semantic structure hinders effective data ingest and harmonisation that would facilitate reusability. As a result, the treasure trove of paleoecological (and other) data is growing without any improvement in ways to allow automated and objective subsequent analysis using e.g. available specialized Python or R libraries and publication of reproducible results i.e. data products or models using Jupyter notebooks or similar environments.

The vision of this proposal is to demonstrate, using abundances of Quaternary planktonic foraminifera as a model, the feasibility of making long-tail data with complex semantics reusable by developing and implementing a community-based workflow covering the full life cycle of such data ranging from data acquisition via data archiving to publication of data products. The task has three aspects: designing a way to sustainably resolve the semantic barriers, implementing the resolved semantics on legacy data and developing a pipeline for submission of new data. The ultimate goal is to design a workflow which can serve as a model that can be adopted in other user communities who need to access and reuse data without incompatibility or semantic barriers, analyse it using e.g. community developed Python or R libraries or dedicated analytical pipelines

and publish resulting advanced and reproducible data products in appropriate archives within NFDI and EOSC.

II. *Pilot description*

The project will rely on the information system PANGAEA (Data Publisher for Earth & Environmental Service) as the technological backbone. PANGAEA is an established long-term archive for the paleoceanography community, hosting a substantial share of the involved publicly available datasets (a search for plankt*+foraminifera+abundance alone returns over 5,000 individual datasets). We will further utilize cloud-based analytical platforms e.g. Jupyterhub serving notebooks (R, Python) either delivered as NFDI core service or alternatively those of EOSC (e.g. EGI notebook). The task of resolving the semantics of taxon names will make use of specialized ontologies such as WORMS (World Register of Marine Species, http://www.marinespecies.org/), made available through NFDI services such as the NFDI4BioDiversity terminology service.

We will find and identify appropriate data via standard search interface and catalogues such as the PANGAEA search engine as well as appropriate catalogue services delivered by NFDI and EOSC. We will investigate how far PROV-O ontology based provenance documentation can be used within data product publication workflows. We will make use of persistent identifiers (DOI) and associated resolver services to identify and access research data as well as to enable data publication and citation of data products.

In order to facilitate reusability of legacy data and stop increasing the number of newly generated orphaned long-tail datasets, the pilot, while focussing on data re-use and publication, will address all parts of the data life cycle. Based on best practises and requirements of the scientific community we will establish standard data archiving as well as tagging workflows in order to evolve existing as well as new datasets into analysis-ready data, thus enabling optimized FAIRness. The aspect of resolving semantic barriers will be backed by learned societies and smaller circles of data producers as well as data scientists. The aspect of generating new workflows will involve key stakeholders, domain experts, data archiving specialists, ontology managers, research data managers as well as data scientists. A series of workshops will initiate this process followed by iterative improvement of workflows to achieve maximum data quality. We will involve domain specialists within editorial workflows of both, PANGAEA as well as authoritative terminologies such as WORMS. Data scientists will use and improve dedicated libraries such as pangaeapy (Python) or pangaear (R) to load these improved data sets into standard R or Python dataframes. Finally, the data will be used to feed standard as well as emerging data analysis pipelines using the tools mentioned above to e.g model past sea surface temperatures. Resulting data products e.g. Jupyter notebooks will be published while maintaining the relationship to original datasets and literature. Data products will be DOI minted and published in PANGAEA including DOI based citation and provenance tracking using DOIs of used data. All workflows and functions developed or established during the pilot shall make use of available or emerging core NFDI services such as processing capabilities and central services such as the

terminology server and common Jupyterhub installations and will in general be embedded or aligned with the overall NFDI infrastructure:

Compared to the status quo, we will establish new, community backed workflows which will guarantee optimal reusability of archived data. The pilot aims at minimising communication gaps between data archiving specialists such as data curators and stewards and the scientific community to establish innovative routines for data product generation as well as publication of resulting data products in a way that can be transferred to other communities facing the challenge of analysing long-tail data with complex semantic structures.

III. *Relevance for the NFDI4Earth*

The pilot approach will be backed by learned societies and smaller circles of data producers, data managers as well as data scientists. It will involve key stakeholders such as domain experts, data archiving specialists, ontology managers as well as data scientists. The principal beneficiaries will be the scientific communities (re-)using long-tail data with complex semantic structure, who will be given tools to generate sustainable ways of combining legacy and new data into analysis-ready formats. The generated standards and workflows will benefit data curators by streamlining submission of new data and infrastructure providers by generating a model of community-driven process to resolve complex ontology.

The pilot uses a specific example to address a common problem for many other geoscientific research communities facing the challenge of heterogeneous data structures or inadequate ontology support. It may further serve as a best practise example on how to integrate the scientific community within the data archiving and publication workflow. While the focus is on the reusability aspect of FAIR, the pilot will address the full life cycle of research data, thus will improve findability, accessibility and interoperability as well. The pilot is closely linked to the information system PANGAEA which is a certified, internationally renowned long-term archive and data publisher. It can be expected that any developments and improvements in the data workflows will engage many users and guide them as best practice solutions. Further, the pilot forms a natural link to NFDI4BioDiversity and will help to improve networking between the two infrastructures.

IV. *Deliverables*

- Designing and implementing a sustainable (long-term) community-driven process to resolve semantic barriers
- Resolving the semantics of Quaternary planktonic foraminifera taxon abundance and occurrence data
- Designing and implementing a workflow to apply the resolved semantics on legacy data
- Developing a pipeline for submission of new data compatible with the resolved semantics

Workflows and functions developed or established during the pilot shall make use of available or emerging core NFDI services such as processing capabilities and central services such as the

terminology server and common Jupyterhub installations and will in general be embedded or aligned with the overall NFDI infrastructure.

V. *Work Plan & Requested funding*

Task	Q1	Q2	Q3	Q4
Resolving semantics, vocabularies				
Data science support				
Development of workflows				
Data product and workflow publication				

We apply for funding of 100% FTE equivalent for 12 months

References

Jonkers, L., Cartapanis, O., Langner, M., McKay, N., Mulitza, S., Strack, A., Kucera, M., 2020. Integrating palaeoclimate time series with rich metadata for uncertainty modelling: strategy and documentation of the PALMOD 130k marine palaeoclimate data synthesis. *Earth System Science Data*, 12(2): 1053–1081.

Jonkers, L., Hillebrand, H., Kucera, M., 2019. Global change drives modern plankton communities away from pre-industrial state. *Nature*, 570: 372–375.

Siccha, M., Kucera, M., 2017. ForCenS, a curated database of planktonic foraminifera census counts in marine surface sediment samples. *Scientific Data* 4: 170109.

Yasuhara, M., D. P. Tittensor, H. Hillebrand, B. Worm, 2017. Combining marine macroecology and palaeoecology in understanding biodiversity: Microfossils as a model. *Biological Reviews,* 92: 199–215.

Yasuhara, M., Wei, Ch.-L., Kucera, M., Costello, M.J., Tittensor, D.P., Kiessling, W., Bonebrake, T.C., Tabor, C., Feng, R., Baselga, A., Kretschmer, K., Kusumoto, B., Kubota, Y., 2020. Past and future decline of tropical pelagic biodiversity. *PNAS*, 117 (23): 12891-12896.