CHRONOSTRATIGRAPHIC API & RDC SERVICES

Johan Renaudie[1], David Lazarus[1]

[1]Museum für Naturkunde, Leibniz-Institut für Evolutions- und Biodiversitätsforschung, Berlin.

Date of submission: 2020-07-28

Abstract

Study of past climatic, environmental and biotic change relies heavily on accurate geologic age estimates interpreted from chronostratigraphic data: geologic age information reported from the sedimentary rocks from which the climate/environmental data came. Data standards to store, APIs to communicate, or cloud-based tools to use chronostratigraphic data do not exist, thus age estimation is done locally, offline; and frequently inconsistent age estimates result. We propose to create chronostratigraphic services for two significant science communities: paleoceanographers and micropaleontologist-biodiversity workers. An API from the Neptune database (micropaleontology and chronostratigraphy) will enable both direct services to other systems, and replication of data into the conceived NFDI Research Data Commons (RDC - Glöckner et al., 2020). RDC services will be created to provide users chronostratigraphic data, age models, and (via PANGAEAs RDC presence) paleoceanographic data; plus tools to create new models and generate ages for environmental/biodiversity datasets.This pilot will bring important offline research communities into a shared common environment for geologic age and chronostratigraphy. Because of the biodiversity data in NSB it will demonstrate NFDI4Earth's link to the biodiversity community, functioning across broad disciplinary boundaries.

I. Introduction

Geochemical, mineralogical and paleontological data from sedimentary samples are important to understanding current and future earth system change. A major challenge are imprecise geologic age estimates, in part caused by the difficulty in locating and synthesizing highly scattered chronostratigraphic data - defined here as the data that determines the geologic age of the samples (numerical value, or text label such as 'Eocene'). Chronostratigraphic data (e.g. bio-, magneto-, geochemo -stratigraphic) must also be referenced to geochronologic data - the standard time scales and radiometric calibrations that they in turn are based on; and for biostratigraphic data, to taxonomic concepts via the rules governing taxonomic names.

Currently, most chronostratigraphic data are in supplementary data archives from publications, in collection databases, in PANGAEA (Germany) or Neotoma (USA). Only a few, isolated, subject-specific databases with explicitly modeled chronostratigraphic content exist: NSB Neptune (Germany; Renaudie et al. 2020), Macrostrat (USA; Peters et al, 2018), Geobiology Database GBDB (China; Xu et al. 2020). All these databases, archives and public repositories use different concepts to store chronostratigraphic data and metadata. Nor is there any exchange standard for chronostratigraphy: current exchange standards (Darwin Core or ABCD) only support static age estimates. Pilot attempts to create exchange standards (e.g. the EFG extension to ABCD) are incomplete and have not been widely deployed.

However this value is not raw data but interpreted data, which depends on how the age was inferred; what biostratigraphic, magnetostratigraphic or isotopic data were used, how these data were calibrated and to which standard timescale (e.g. a geomagnetic polarity time scale - GPTS) the calibrations were made. As calibrations and standard scales evolve with time, those static age assignments attached to samples are frequently out of date and incorrect. The relevant chronostratigraphic data needed to correct these ages, despite often being in existence, is typically stored in another system and is not linked to the sample records. The absence of linked chronostratigraphic metadata accompanying these age assignments, or a standard for creating the linkage, is a major hinderance to correct data synthesis.

We propose a pilot study which will serve as a seed project towards the ultimate goal of a fully online, commons based standard for storing and using chronostratigraphic data and age estimates from these; and for shared analytic tools. Our vision is the ability for researchers worldwide to rapidly place paleoclimate, paleoenvironmental and paleobiological data from globally networked data sources into a consistent, accurate geologic age framework, as the basis for improved study of past climate and biotic change.

II. Pilot description

Our pilot study will replicate NSB data into an enhanced chronostratigraphic structure into the RDC cloud via a new API. There we will create functions for users to simply apply existing age models to generate geologic age estimates from NSB or PANGAEA to paleoclimate/paleoenvironmental/paleobiodiversity datasets. Further, our new services will provide users with the underlying chronostratigraphic data from these systems, plus integrative tools to generate new age estimates via modified age models for sections.

The Neptune Database (Renaudie et al. 2020; http://nsb-mfn-berlin.de) of fossil occurrences from deep-sea drilling boreholes includes a taxonomic backbone linked to the micropaleontology community's main taxonomic catalog Mikrotax (Young et al. 2019). The user community however consists not only of paleobiodiversity researchers but also paleoceanographers and paleoclimatologists, as the database also contains both a large set of curated age models for these deep-sea sites, and the (mostly biostratigraphic) chronostratigraphic data used to create them. The chronostratigraphic data model developed for NSB uniquely links numerical age models for a given site with the bio, magneto and other stratigraphic events observed from the site, their calibrations, the geomagnetic polarity time scale (GPTS) used for the calibration and metadata for all these, while supporting multiple interpretations for each of them.

PANGAEA is a very large archive of published earth science data and part of the World Data System of the ISC. The primary curation object is a dataset, which is retrieved largely in the format submitted, together with rich metadata headers. PANGAEA has become a primary online archive for earth science data from deep sea sections, and in particular the deep sea drilling programs cored sections that are the source of chronostratigraphic data for NSB. The MfN has long established communication links with PANGAEA staff and specifically has discussed the pilot study proposed here.

In addition to NSB and PANGAEA, there are many existing sources for our project, e. g. more generalized data models for geologic sections (Macrostrat, Neotoma); prototype APIs in the EFG extension to ABCD and the EarthLife Consortium; and integration tools in our own apps, e.g. ADP and its SOD chronostratigraphic data format (Renaudie et al., 2020). Our task will be 1) to migrate these concepts into the RDC environment; 2) create the replication system from NSB via an API; 3) create the new services integrating age models and chronostratigraphy from NSB with PANGAEA data. As the RDC is still conceptual, the essential steps can only be briefly listed. Software tools and documentation standards are detailed in Glöckner et al. (2020).

Task 1 - creating a chronostratigraphic data backbone in the RDC. Current data models (cited above) are implemented in relational table structures in postgreSQL and we will initially create our generalized model by merging and extending them in this environment, where the populated data systems currently exist and the correctness of the content can be quickly validated. We will also create extensions for currently non-modelled characteristics e.g. composite depth models, astrochronology and isotope trend curves. This structure will then be mapped-moved to the backbone in the RDC environment (which may be a non-sql based system).

Task 2 - create an API and a replication system for the chronostratigraphic content of NSB in the RDC. Here we can make use of the MfN's FB3 data management group's expertise in creating Restful APIs, NSB's data model for age models, events and linked chronostratigraphic data types, the MfN co-developed -EFG extension to ABCD for non-deep sea drilling data sources, plus concepts for discontinuous sections from Macrostrat.

Task 3 - create the user services in the RDC. We have in prior work created basic data presentation/integration code for NSB's website, scripting packages for data integration between NSB and PANGAEA (Renaudie et al, 2020) including age estimates for PANGAEA data from NSB age models. Basic age estimation services can largely use existing code as a template. The more extensive interactive graphic app ADP will be the code template to present chronostratigraphic and age model data, together with PANGAEA datasets in a data rich RDC based user service. NSB's website and ADP are both coded in Python and thus, assuming Python will be supported in the RDC environment, good fits for e.g. Jupyter notebooks. Indeed, while the I/O portions of the code will change, many internal blocks may be reusable with only minor modification.

III. Relevance for NFDI4Earth

The API and enhanced chronostratigraphic data model can be used in many contexts, e.g. to extend, and improve direct communication between, external databases (GBDB, Macrostrat etc.) The expected users of our RDC services will include hundreds of Germany/EU-based paleoceanographers, paleoclimatologists, paleobiodiversity researchers plus other scientists who need to place their data into a geologic age framework for synthesis, but who currently are working offline and in isolation. By making data easily accessible, and by not requiring highly specialized expertise to locate and synthesize disparate chronostratigraphic data sources (e.g. biostratigraphic data), the services should improve age estimate quality, and the research based on them.

Our proposal supports FAIR. It improves Interoperability by creating a common cloud-based data commons to communicate between various geological data sources, and online tools for users to integrate into their own workflows. The final product will however also be a significant milestone in addressing the accessibility and reusability aspects of FAIR as it will allow us not only to remobilize data from unstructured, non-standardized data repositories into structured, RDC accessible datasets but also offer a set of common-ground metadata that will allow the users to understand and use appropriately the geologic age estimates derived from these data. Lastly this project will link biodiversity databases, with NSB as an example, to earth science data, and thus demonstrate the ability of NFDI to function across broad disciplinary boundaries.

IV. Deliverables

The first deliverable is a data structure (first implemented in NSB as SQL tables, but then implemented in the NFDI-RDC) that is able to not only record appropriately the chronostratigraphy of deep-sea sedimentary record but also (in order of importance) the chronostratigraphy of land sections of marine sediments, of terrestrial sediments and single-objects. This data structure needs to also be able to abstract calibrations of non-discrete chronostratigraphic data, i. e. astrochronology, isotope stratigraphy and linked data from geochronology. Additional metadata may also need to be gathered from the literature or from data repositories to populate this structure (in particular the fields not currently implemented in NSB).

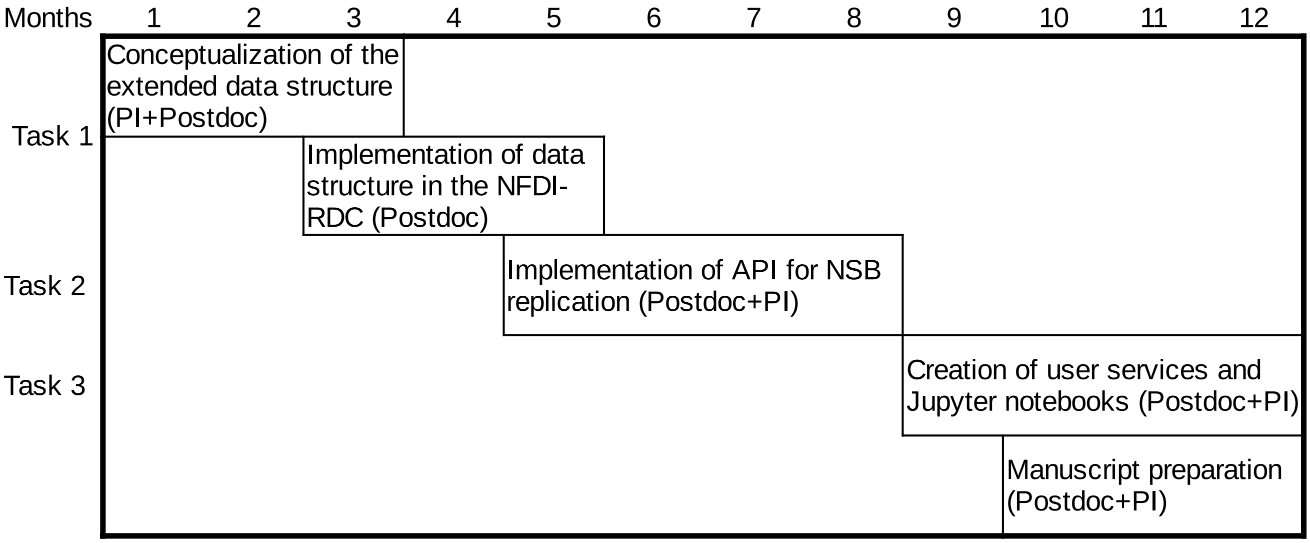
The second deliverable is an API based on the first deliverable that can be used to retrieve chronostratigraphic data from the newly extended NSB, and exchange such data between other deeptime databases, and replicate NSB into the NFDI-RDC.

Additionally, python-based user services will be developed and made accessible in the RDC as e.g. Jupyter notebooks allowing users to compute sample ages for given datasets and access the underlying stratigraphic metadata for reproducibility purposes; to explore age models using an interface similar to that developed for the AD software (see Renaudie et al. 2020); to access state-of-the-art calibration data for the user to produce new age models not present yet in the system.

All deliverables will be made available using creative commons licences, and a publication describing the results will be created.

To create our roadmap: Chronostratigraphic Data Standards and Services, we will (virtually) present at meetings, and host workshops involving major database system managers (e.g. as cited above) and representative users.

V. Work plan & Requested funding


Months
1
2
3
4
5
6
7
8
9
10
11
12
Task 1
Task 2
Task 3
Conceptualization of the extended data structure (PI+Postdoc)
Implementation of data structure in the NFDI-RDC (Postdoc)
Implementation of API for NSB replication (Postdoc+PI)
Creation of user services and Jupyter notebooks (Postdoc+PI)
Manuscript preparation (Postdoc+PI)

We are requesting one year of funding for a postdoc (preferably a geoscientist with a strong computer science background) who will take the lead on the implementation aspect of the project (data structure, API, services, manuscript).

References

Glöckner, F. O., Diepenbroek, M., Felden, J., Güntsch, A., Stoye, J., et al. (2020, July 14). Zenodo, http://doi.org/10.5281/zenodo.3943645
Peters, S. E., Husson, J. M., Czaplewski, J. (2018). *Geochem.Geophys.Geosyst.*, 19(4):1393-1409.
Petersen, M., Glöckler, F., Kiessling, W., Döring, M., Fichtmüller, D., et al. (2018) *Foss.Rec.*, 21:47-53.
Renaudie, J., Lazarus, D.B., Diver, P. (2020). *Pal.Electron.*,23(1):a11.
Xu, H.-H., Niu, Z.-B., Chen, Y.-S. (2020) EarthSyst.Sci.Dat.Disc. https://doi.org/10.5194/essd-2020-164.
Young, J.R., Bown, P.R., Wade, B.S., Pedder, B.E., Huber, B.T., et al., (2019). Act.Geol.Sin., 93:70–72.