AGU100 ADVANCING EARTH AND SPACE SCIENCE

Geochemistry, Geophysics, Geosystems

TECHNICAL REPORTS: METHODS

10.1029/2018GC007467

Macrostrat: A Platform for Geological Data Integration and Deep-Time Earth Crust Research

Shanan E. Peters[1], Jon M. Husson[1,2], and John Czaplewski[1]

[1]Department of Geoscience, University of Wisconsin-Madison, Madison, WI, USA, [2]School of Earth and Ocean Sciences, University of Victoria, Victoria, BC, Canada

Key Points:

- Macrostrat is a geospatial database containing comprehensive information on bedrock geology in the surface and subsurface
- Geospatially resolved ages, lithologies, and attributes of the upper crust facilitate quantitative analyses and data integration
- Macrostrat data are accessible via an Application Programing Interface invokable from within scientific, mobile, and web applications

Correspondence to:
S. E. Peters,
peters@geology.wisc.edu

Citation:
Peters, S. E., Husson, J. M., & Czaplewski, J. (2018). Macrostrat: A platform for geological data integration and deep-time earth crust research. *Geochemistry, Geophysics, Geosystems*, *19*. https://doi.org/10.1029/2018GC007467

Received 1 FEB 2018
Accepted 26 MAR 2018
Accepted article online 16 APR 2018

Abstract Characterizing the lithology, age, and physical-chemical properties of rocks and sediments in the Earth's upper crust is necessary to fully assess energy, water, and mineral resources and to address many fundamental questions. Although a large number of geological maps, regional geological syntheses, and sample-based measurements have been produced, there is no openly available database that integrates rock record-derived data, while also facilitating large-scale, quantitative characterization of the volume, age, and material properties of the upper crust. Here we describe Macrostrat, a relational geospatial database and supporting cyberinfrastructure that is designed to enable quantitative spatial and geochronological analyses of the entire assemblage of surface and subsurface sedimentary, igneous, and metamorphic rocks. Macrostrat contains general, comprehensive summaries of the age and properties of 33,903 lithologically and chronologically defined geological units distributed across 1,474 regions in North and South America, the Caribbean, New Zealand, and the deep sea. Sample-derived data, including fossil occurrences in the Paleobiology Database, more than 180,000 geochemical and outcrop-derived measurements, and more than 2.3 million bedrock geologic map units from over 200 map sources, are linked to specific Macrostrat units and/or lithologies. Macrostrat has generated numerous quantitative results and its infrastructure is used as a data platform in several independently developed mobile applications. It is necessary to expand geographic coverage and to refine age models and material properties to arrive at a more precise characterization of the upper crust globally and test fundamental hypotheses about the long-term evolution of Earth systems.

1. Introduction

Alexander Ronov's group, at the Vernadsky Institute of the Russian Academy of Sciences, were among the first geoscientists to demonstrate the scientific value of compiling spatially and temporally comprehensive data on the age, lithology, and volume of rocks in the Earth's crust (Ronov, 1982, 1994; Ronov et al., 1980; Ronov & Khain, 1954; Ronov & Migdisov, 1971; Ronov & Yaroshevsky, 1969). Using a combination of geological maps and borehole observations, Ronov and his team generated global rock volume and chemical composition estimates for general lithology types across geological epochs (or longer duration time intervals) in the Phanerozoic and latest Precambrian. In addition to providing the first data with which to quantitatively describe the rock record, Ronov's compilation served as a basis for direct estimates of temporal changes in burial and weathering fluxes of biogeochemically important elements (e.g., Berner, 1989; Budyko et al., 1987; Wilkinson & Walker, 1989) and for constraining a wide range of quantities, ranging from groundwater volumes (e.g., Gleeson et al., 2016; Hay & Leslie, 1990) to rock cycling rates (e.g., Gombosi & Wilkinson, 2012; Wilkinson et al., 2009; Wilkinson & Walker, 1989; Wold & Hay, 1990).

Although Ronov's work was a scientific success and played a key role in the development of some of the first models describing the geochemical evolution of Earth's surface environment (Berner & Canfield, 1989; Budyko et al., 1987), his team's initial compilations were inherently low resolution. The reason stems from the pioneering nature of the work, most of which was carried out in the 1950s through 1970s, and from a dependence on small-scale geologic maps and limited borehole data. Both types of records focus on physical contact relationships and the spatial extent of general bedrock types. However, whatever Ronov's compilation may have lacked in temporal and lithological acuity was, in many ways, compensated for by the fact that it was globally comprehensive. The subsequent emergence of isotope-based approaches to

deciphering changes in Earth systems (e.g., Des Marais et al., 1992) shifted emphasis away from Ronov's laborious approach of compiling data on the rock record and toward the production of new geochemical proxy records, which could be extracted with more efficiency and with higher temporal resolution in one or more well-correlated stratigraphic sections. Nevertheless, most of the models that are used to interpret geochemical proxy records require that assumptions be made about burial and weathering fluxes, which are difficult to assess without independent data on the rock record (Bergman et al., 2004; Berner & Kothavala, 2001; Halevy et al., 2012; Husson & Peters, 2017; Schrag et al., 2013). Thus, there remains a need for spatially and temporally complete quantitative descriptions of the rock record that can be combined with geochemical models and other proxy records. Data on the rock record are also critical to calibrating physical models of the upper crust that can be used in modeling ground water volume (e.g., Gleeson et al., 2016), fluid flow (e.g., Fan et al., 2016), and geophysical heterogeneties (e.g., Mooney et al., 1998).

There are several approaches that could be taken to arrive at a comprehensive space-time description of the upper crust that is useful for both scientific questions and for informatic initiatives. One end-member approach aims for the highest-possible resolution and uses only vetted, authoritative primary field data. This methodology is most useful when constructing databases that are targeted for specific measurement types, when effort is focused on small geographic regions, or when it is necessary to maximize precision and accuracy in order to address questions that depend critically on individual observations (e.g., the oldest fossil of a given animal clade, Benton et al., 2015). However, restricting data to only what is considered today to be the most up-to-date and/or best available is impractical when characterizing the entire rock record on a continental or global scale. By definition, the best and highest-resolution data either have not been collected yet or are sparse relative to the full extent of the upper crust. An alternative approach is, therefore, to start with basic geological summaries that are spatially and temporally complete, but that may lack the highest-possible resolution. Such comparatively simple geological summaries are based ultimately on primary field data and observations, but primary data are not the focus.

Macrostrat's main objective is to aggregate and systematize basic field-derived data products, such as geological maps and regional geologic columns, in order to synthesize a large number of primary field observations and measurements into a spatially and temporally complete description of the upper crust that can be enhanced with new data and information. There are several reasons why this starting point is useful, both scientifically and from an informatics perspective. For example, low-resolution, but also temporally and spatially complete, descriptions of the basic space-time attributes of the rock record serve as a useful basis for estimating rock quantities and ages (in the sense of Ronov) and for assessing the stratigraphic distribution of proxy data, such as the stratigraphic completeness of paleontological sampling (Peters & Heim, 2010). In addition, general but complete summaries of the upper crust can be used to link geological data sets and to constrain the ages of their constituent rock record-derived data, in both a relative and absolute sense.

Here we describe the motivation for building Macrostrat and its general data model. Although the database has focused on comprehensive large-scale data, its fundamental architecture is scale independent and can accommodate the highest-resolution field data. We then outline the cyberinfrastructure that currently supports the database and describe how that infrastructure can be accessed by software via an Application Programing Interface (API). The API currently supplies basic geological data to several third-party applications built to support field work, data analysis, and educational and outreach activities. One such application is outlined as a working example. Finally, we present a general overview of the data currently in Macrostrat and provide some basic results that demonstrate Macrostrat's scientific utility while at the same time exposing the need for extending geographic coverage globally.

2. The Macrostrat Data Model

Macrostrat is a relational geospatial database deployed on a unix-based systems in both a MariaDB and PostGIS-enabled PostgreSQL environment. The database is designed primarily to facilitate quantitative macrostratigraphic analysis of the entire upper crust. Macrostratigraphy is an analytical approach that is inherently chronostratigraphic in nature (Peters, 2006). The basic unit of analysis is, therefore, a temporal gap-bound package of rock identified at a single geographic location (Hannisdal & Peters, 2010; Peters, 2008a). A rock package can consist of any lithology, but the gaps that define the boundaries between packages

depend on the operational definition of a "gap." For example, if a gap is defined as a break in temporal continuity (e.g., a hiatus), then it is necessary to specify a duration threshold for gap recognition (e.g., 1 Myr). This gap-duration threshold then renders the continuum of temporal continuity that is inherent in the rock record into a binary distribution (presence and absence of rock of a given age at a given location). Alternatively, a gap could be defined by lithological attributes (e.g., a temporal gap in siliciclastic sediments could be occupied by a hiatus or by a shift to carbonate sediments). The analytical approach of macrostratigraphy is scale independent, and the ideal data set is compiled at the finest possible spatial, temporal and lithological resolution so as to allow the application of any arbitrary gap recognition criteria and scale of analysis (e.g., Aswasereelert et al., 2013). However, the strength of the current Macrostrat data set is its ability to characterize spatial and temporal variation that occurs on the scale of a basin, geological province, or continent.

The quantitative framework defined by macrostratigraphy is a good descriptor of the core organizational concept that motivated database development, but the design of the database is organized around an even more basic component and includes several additional features, which are outlined below. Note that Macrostrat column data include igneous, metamorphic and sedimentary rocks, but for clarity, the discussion herein will focus largely on sediments, which are present in 90% of the 33,931 total rock unit records currently in the database.

2.1. Macrostrat Units

The fundamental object in the Macrostrat database is called a "unit" and it represents a body of rock or sediment that is recognized at the time of data compilation as being genetically, lithologically and/or chronologically distinct from other such adjacent units. A Macrostrat unit could, therefore, consist of a thickness of sediment identified in a measured section or core (e.g., a bed), or it could consist of a lithostratigraphic formation or other rock body that is described as physically and temporally distinct in a regionally composited geological record (e.g., a geological map unit or a lithostratigraphic unit in a regional stratigraphic column). In all cases, Macrostrat units are recognized separately within each geographic region, referred to as a "column." Each Macrostrat "column" consists of a geospatial footprint (polygon) defining the lateral boundaries of the focal region, along with descriptive metadata for that area, such as references supplying column data, column name, etc. (Figure 1).

Units in Macrostrat are described by a variety of physical attributes, including thicknesses (usually expressed as a maximum and a minimum within the focal region), one or more lithologies, attributes that modify lithologies, inferred environments of deposition/emplacement, and stratigraphic nomenclature (Figure 1). All of the attributes that can be assigned to Macrostrat units, and the lithologies that they contain, are stored as dictionaries in separately managed database tables that also contain related information, such as hierarchy (e.g., sedimentary rock includes carbonate, carbonate includes grainstone) and synonyms (e.g., "dolostone" is an alternative name for "dolomite").

All Macrostrat units have at least one dominant lithology assigned to them, but multiple lithologies, and the relative volumetric abundances of each, can be recorded for each unit, the latter either quantitatively or semi-quantitatively (i.e., either numerical estimates of proportional abundance can be supplied or qualitative statements of abundance, "dominant" and "subordinate," can be applied to each lithology; the latter are computed into proportional estimates using simple rules). One way of improving existing Macrostrat data is to refine the lithological information that is linked to units, either by providing additional lithologies or by adding more detailed attributes to lithologies (e.g., "stromatolitic dolostone" rather than "dolostone").

Each Macrostrat unit is treated as a distinct entity that is associated with only one column, regardless of whether or not the geological object that a unit represents can be traced laterally between multiple adjacent columns. Columns are, therefore, equivalent to independent samples of the upper crust. This design allows spatial variability in the attributes that are assigned to "units" to be captured. For example, a widespread time-transgressive lithostratigraphic rock unit would be intersected by multiple Macrostrat columns and the age of the corresponding units assigned to that lithostratigraphic unit in each column could be different, reflecting its time transgressive nature. In this case, only the lithostratigraphic name applied to the units would identify them as related in some way (thereby demonstrating the poor time-correlation value of that particular lithostratigraphic name). Similarly, a single rock body that varied spatially in thickness and lithology would be represented by multiple units with different such attributes in each intersecting column.

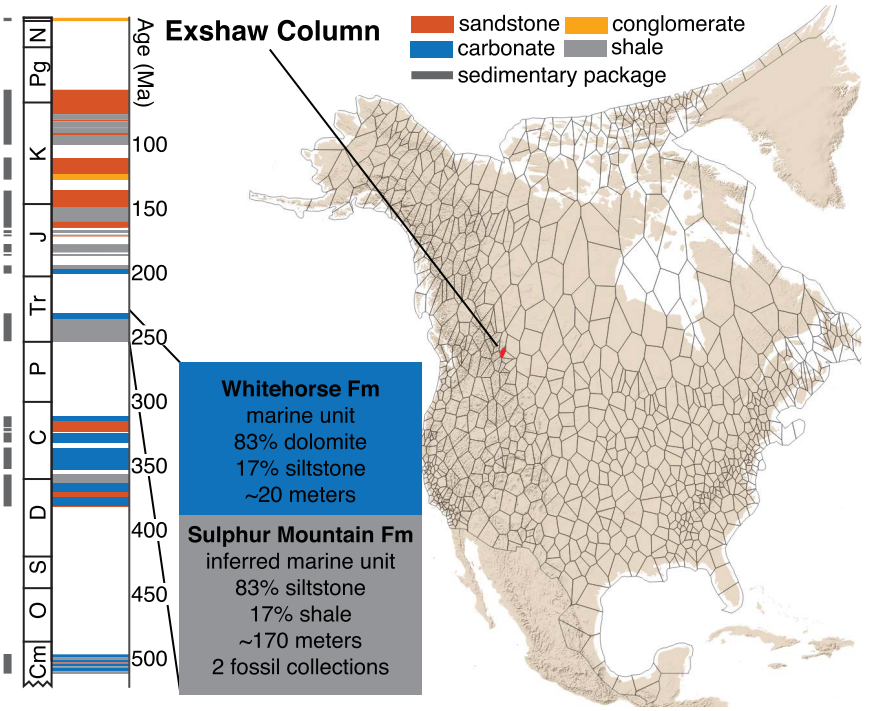

Figure 1. Macrostrat North America, with an example column (titled "Exshaw") highlighted in red on the map and rendered chronostratigraphically in the column on the left. Each unit is colored by its dominant lithology and grouped into packages on the basis of temporal continuity (indicated by the solid dark bars left of the geologic time scale). Although only the dominant lithology is represented by the colors of units, many units include multiple lithologies and estimates for the relative volumetric abundance of each (e.g., the expanded Triassic units identified as the Whitehorse Formation and the Sulphur Mountain Formation consist primarily, but not exclusively, of dolomite and siltstone, respectively).

The primary grouping criteria for Macrostrat units is their assigned geographically defined column, but units within a column are also grouped into "packages" (also known as "sections") on the basis of temporal and genetic continuity that is defined at the time of data entry. Package structure can also be calculated dynamically depending on the criteria that are used to define units of interest and the gaps that separate them (see above). An advantageous approach, then, is to create columns with the finest possible temporal and lithological resolution, which allows packages to be defined and analyses to be conducted at any arbitrary scale.

Dictionaries of known terms, including lithologies, attributes that modify lithologies (e.g., "bioturbated sandstones"), environments of formation (e.g., "shoreface"), minerals, measurements (major element chemistry), chronostratigraphic time intervals, and lithostratigraphic names are stored in Macrostrat, but there is no attempt to be prescriptive about how these terms are applied to individual units. Doing so would effectively prohibit the use of a large fraction of the published primary field descriptions and data. Ambiguity, uncertainty, and inaccuracy are, therefore, expected in some cases. For example, the lithology "marl" is a mixture of terrigenous clay and fine carbonate sediment, and that lithology can be assigned to a Macrostrat unit, without any indication of the relative mixtures of the two components in "marl." However, insofar as descriptions of rocks and the data extracted from them can be geographically located and linked to a physically recognizable rock body, Macrostrat can help to organize observations and provide constraints on properties that are commensurate with the precision of the language used to describe rocks. The overarching goal is to arrive at a working and inclusive description of the upper crust that can be refined as newer and better information and data are incorporated.

2.2. Macrostrat Columns

The geographic columns that contain Macrostrat units are of two basic types: (1) those that represent a precisely located and discrete sample of the upper crust, like those supplied by measured sections or boreholes in offshore drilling sites from the Ocean Drilling Program, Deep Sea Drilling Project, and International Ocean

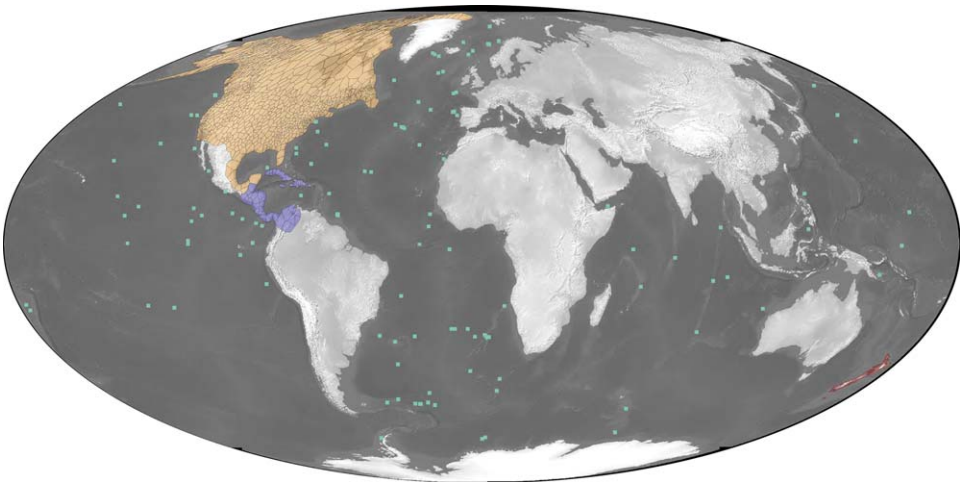

Figure 2. Geographic distribution of columns, segregated by project (North America, Caribbean, New Zealand, and deep sea), in the current public version of the Macrostrat database. Columns located on continental crust acquire, by default, a geographic footprint that is defined by a Voronoi tessellation. The points used to create the tessellation correspond to the approximate center of the region covered by each composite column. It is possible to edit the spatial topology of columns in order to align their boundaries with geologically meaningful features. Macrostrat columns from the deep sea are assigned point coordinates based on the offshore location of each drilling site. For consistency with continental columns, offshore drill sites are represented by rectangular buffers around those points.

Discovery Program and (2) those that represent a composited summary of the geology over a geographic area (e.g., Childs, 1985; King et al., 1999; Maurrasse, 1990; Stott, 1991; Trettin, 1991). Both data types have advantages and disadvantages. For example, composited geological columns sacrifice spatial information within the region they cover, but as a result they can capture units with limited geographic extents that would likely be missed in single boreholes or measured sections.

Macrostrat currently consists of four major groups of columns that are separated, for convenience, by geographic region (Figure 2). Columns are assigned to "projects" that identify these groups, some of which might share primary source reference(s), compilation approaches, or regions. For example, the deep sea data set consists entirely of core descriptions compiled from offshore drilling sites (Peters et al., 2013; Fraass et al. 2015), whereas the continental record in North America consists of regionally composited geologic columns. Because the latter typically lack precise definitions of geographic extent, the boundaries between all such composited columns in Macrostrat are currently interpolated using Voronoi tesselation and a manually constructed bounding geometry. The boundaries of Macrostrat column polygons could be modified to reflect actual geological and structural boundaries and other geographic and geologic features, but doing so requires additional geospatial data to define relevant boundaries and/or manual effort to adjust the tesselation boundaries.

2.3. Geochronological Time Intervals

Chronostratigraphic time intervals (e.g., biozones, ages, and epochs) are stored in Macrostrat and related to one another and to numerical ages in both relative and absolute senses. Chronostratigraphic intervals that have actual numerical age estimates, principally those provided by the International Commission on Stratigraphy for Global Stratotype Boundary Section and Points (GSSPs; Gradstein et al., 2012), are referenced to absolute time (subject to explicit uncertainties and future revision). Chronostratigraphic intervals for which there are no direct numerical age constraints are not assigned numerical ages. Instead, intervals lacking direct geochronological constraints are assigned boundaries with positions that are defined relative to another chronostratigraphic interval (e.g., a boundary for a chronostratigraphic bin could be referenced to $25 \pm 5\%$ of the duration through an international age, which is in turn referenced to boundaries that do have absolute numerical age estimates, such as GSSPs). This approach to managing geochronological time intervals and their numerical ages obviates the need to associate each interval with an explicit stored numerical age(s). It also makes actual age constraints more transparent and has a data management advantage.

Chronostratigraphic time scales (e.g., international ages and periods, biozonations, regional chronostratigraphic subdivisions) and reference information for each time scale are stored in Macrostrat. However, because a chronostratigraphic time scale is essentially a group of individual named time intervals, time scales are only indirectly referenced to intervals (in much the same way that units are only indirectly referenced to columns via a join in the database). This approach allows a one-to-many relationship between time intervals and the time scales that use them (i.e., the Rhaetian is an international age, as well as part of the North American regional time scale), which in turn enables the creation of custom time scales from existing time intervals.

2.4. Continuous-Time Age Model

Each Macrostrat unit that is not directly associated with a geochronological measurement (e.g., a radiometrically dated ash bed) acquires an initial modeled numerical age by applying basic correlation approaches and by using temporal contact relationships with other units in the same column to constrain the distribution of time through rock volume. Because the units that comprise Macrostrat columns are often shorter in duration than the chronostratigraphic time bins to which they can be correlated, basic laws of superposition and relative aging allow time to be distributed more finely within and between units than bin-based correlations. This means, for example, that Macrostrat's continuous-time age model is capable of predicting the age of an ash bed (with error determined by positional uncertainty) before the relevant measurement is made. In the case of a numerically dated unit, its boundaries are referenced to an absolute position in time, with analytical uncertainty. The latter chronological "spikes" anchor the position of the bed in time and can serve as constraints in an incrementally improving age model.

Although capable of incorporating direct numerical age estimates, the preliminary age model for Macrostrat was constructed for each column using only the chronostratigraphic bins to which its constituent units are correlated and the relative temporal constraints provided by contact relationships. For example, if there were 10 vertically stacked sedimentary units assigned to one continuous package (units are always assigned to only one column), and if together those units spanned completely one chronostratigraphic time bin (e.g., the Frasnian), then the absolute time represented by that chronostratigraphic bin would be distributed equally and sequentially between each successive unit. The oldest unit would have a bottom age equal to the base of the chronostratigraphic bin (i.e., 0% of the way through the Frasnian), the youngest unit would have a top age equal to the top of the chronostratigraphic bin (i.e., 100% of the duration through the Frasnian). Unit boundaries between these two stage-defining tie points would be distributed equally and proportionally within the chronostratigraphic bin (e.g., the top of the first unit/bottom of second unit would be at position equal to 10% of the duration of the Frasnian, the top of the second unit/bottom of the third unit would be at a position equal to 20% of the duration of the Frasnian, etc.). Because many Macrostrat columns are regionally composited, it is not uncommon for there to be coeval units in a single column (i.e., there are "laterally adjacent" units that reflect spatial variation in lithology within the geographic region covered by the column). In such cases, units in the same column will have overlapping ages in the age model. Note that physical contact relationships are not reflected in this model. For example, a dike would have chronostratigraphic boundaries with other units in time, but might physically cut across all units in a given column. To address the latter contact relationships, geological map-type data are required (discussed below).

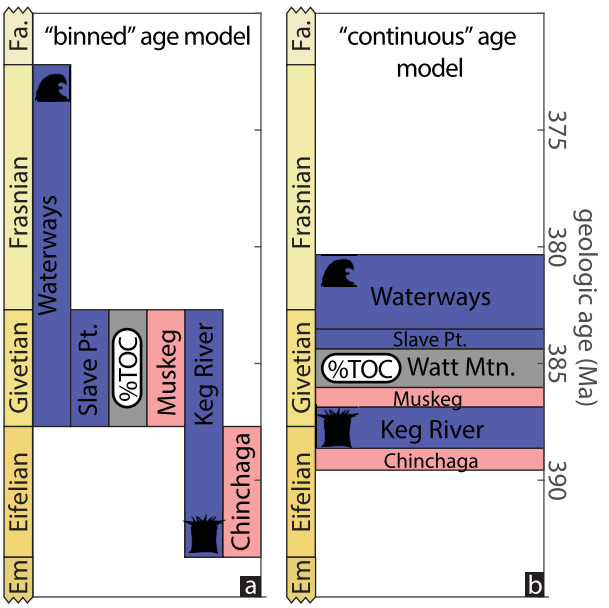

Figure 3. Illustration of (a) "binned" versus (b) "continuous" age model, using a Devonian gap-bound package from the Zama Lake column in northern Canada. In this example, units were originally correlated to one or more epochs (Figure 3a). Using superposition (i.e., Chinchaga is overlain by Keg River) and more refined opinions about the correlation of units to a chronostratigraphic time intervals (i.e., the top of the Waterways Formation is found in the lower half of the Frasnian), the result is an internally consistent continuous age model (Figure 3b). The boundaries of units have identity and serve as the basis for age assignments in Macrostrat. Units with matched Paleobiology Database fossil collections have icons of an example taxon selected from "prevalent taxa" based on occurrence counts.

Macrostrat's initial age model uses the fewest possible parameterizations (i.e., correlation to chronostratigraphic time intervals and superposition) to arrive at an internally consistent continuous-time age model (Figure 3). As a result, the model lacks principled statements of uncertainty and does not take advantage of all other information that

could be temporally informative, such as thickness and lithology. However, the age model is readily improved. Data produced by geochronological laboratory facilities, for example, can be incorporated into Macrostrat's age model and then propagated to all data resources linked to it (e.g., a radiometrically dated ash bed within a Macrostrat unit would automatically constrain the ages of all fossil and geochemical samples linked to that unit and adjacent units, as in Figure 3). Additional approaches to producing age models, including, for example, constraints from event ordination (Sadler, 1981) and astrochronological tuning (e.g., Meyers et al., 2012) can also be incorporated.

By integrating numerical age estimates into a comprehensive framework describing the rock record, geochronological lab facilities can readily acquire broader geological context for measurements and help prioritize effort by identifying geographic or temporal segments of the rock record that could benefit the most from new measurements. Because the unit-based architecture of Macrostrat can accommodate any arbitrary scale of subdivision of rock, incorporating a new dated horizon in a sedimentary unit requires only dividing the containing unit into the dated horizon and adjacent components (e.g., a dated ash bed from the middle of a single Macrostrat unit would require division of that unit into three portions, the portion of the unit below the bed, the bed itself, and the portion of the unit above the bed). The new dated horizon would then serve as a "spike" constraint on numerical age, with error defined by the analytical precision of the measurement and the positional uncertainty of the bed within the unit.

2.5. Lithostratigraphic Names and Hierarchies

Macrostrat manages the names that are assigned to rock units (e.g., lithostratigraphic members and formations) in three ways. First, "concepts" are used to designate groups of names that identify the same entity. For example, the "Dakota" concept applies to lithostratigraphic names of formation rank, including the "Dakota Sandstone," "Dakota Formation," and "Dakota Conglomerate." The stratigraphic concept of "Dakota" also applies to a lithostratigraphic name of group rank, the "Dakota Group." All four of these lithostratigraphic names and ranks are separately stored in Macrostrat but they are also all identified as belonging to the same lithostratigraphic concept: "Dakota." Concepts are also associated with additional information, including descriptions of usage, geologic age, general lithological and/or temporal properties, geographic region, and source reference. The overall structure of the concept component of Macrostrat's lithostratigraphic nomenclature is comparable to the USGS Lexicon (USGS, 2016).

In addition to grouping lithostratigraphic and other rock-body names that refer to the same geological entity, Macrostrat explicitly stores nomenclatural hierarchy. For example, the "Dakota Formation" (one of the names and ranks used in the "Dakota" concept) is the parent of four member-level lithostratigraphic names. Explicit storage of nomenclatural hierarchy makes it possible to access Macrostrat data from any nomenclatural starting point and to then obtain all of the parent and child lithostratigraphic names and their variants, as well as the rock units to which they are applied in space and time.

Currently, more than 36,000 lithostratigraphic names are stored in Macrostrat, most of which derive from modified versions of the USGS National Geologic Map Database, Australian Lexicon, Canadian Weblex, and British Geological Survey Lexicon stratigraphic lexicons, as well as other external resources. Reference to these sources and URLs linking back to original Lexicon data pages are provided for concepts wherever possible, but most of the relevant information associated with stratigraphic names is also available from within Macrostrat.

Lithostratigraphic names are notorious for lacking chronostratigraphic significance and, in some cases, for dizzying historical convolutions. However, this fact does not diminish their prevalence in the published literature or their usage on geologic maps, the field books of geologists, and museum specimen labels. Lithostratigraphic names are, in many regions of the world, the lingua franca for parts of the rock record that are, at least in principle, readily recognizable in the field. Macrostrat's data structure is capable of storing lithostratigraphic terms and doing so in a way that exposes their spatiotemporal disparities and inconsistencies. Indeed, the ability of Macrostrat to provide a quantitative space-time index of lithostratigraphic nomenclature is one of the informatics-related strengths of the database (see below).

Like most components of Macrostrat, there remain ambiguities and errors in the nomenclatural hierarchy and the assignment of names to Macrostrat units. For example, it is possible for some lithostratigraphic homonyms to not be resolved. Such ambiguities are readily fixed when they are discovered, and any

changes made to the database propagate automatically. Just as the field of geology (and all empirically grounded science) remains in a constant state of refinement and improvement, none of the information in Macrostrat should be viewed as static. The database continues to improve as human expertise is applied to the process of data curation and as new constraints on and hypotheses for the chronology and physical properties of the upper crust emerge.

2.6. Geologic Maps

Bedrock geologic maps are working hypotheses for the surface expression of physical, three-dimensional rock bodies and structures in the upper crust. Maps are typically derived from a combination of aerial imagery and field-based measurements and observations, which are then transformed into spatially complete models using widely accepted (but heterogeneously applied) methods and criteria (Chorlton, 2007; Garrity & Soller, 2009; Raymond et al., 2012). Similar to Macrostrat columns, which constitute working hypotheses for the chronological distribution of rock bodies that can be refined by the addition of new constraints, new field data and observations can result in revisions of a geologic map. Maps (and Macrostrat columns) are, therefore, more akin to model output than they are to primary data. Nevertheless, geologic maps are useful starting points for framing geological field problems and for motivating additional data collection and hypotheses. They can also serve as useful data in their own right (Peters & Husson, 2017; Raup, 1976; Smith & McGowan, 2007; Wall et al., 2009; Wilkinson et al., 2009). Many of the publications and unit descriptions accompanying geological maps also contain detailed field descriptions of rock units that are often underutilized. One objective of the geologic map component of Macrostrat is to expose the information behind geological maps to a wider range of uses, ranging from the facilitation of geological field work to data synthesis tasks.

All bedrock and surficial geologic maps consist fundamentally of geospatial polygons and, optionally, lines and points, all with associated attributes. Polygons represent geologic map units, believed by the authors to have some physical and/or genetic continuity. Lines represent faults, fold axes, dikes, marker beds, and other surface-expressed lineaments. Points describe the location of fabric orientation measurements (e.g., foliation and bedding strike-dip), and other measurements (e.g., paleocurrent directions) and location-specific observations (e.g., mineral/fossil occurrences). Macrostrat's PostGIS geological map database stores three groups of information for all bedrock and surficial geological maps: (1) the original vector-based map objects (polygons, lines, and points) and their attributes, transformed into the PostGIS environment, (2) standardized representations of maps that include elements common to all geological map objects (see below), and (3) tables that store intersections of geological map objects and Macrostrat entities (i.e., units, lithologies, lithostratigraphic names, and chronostratigraphic intervals).

The original sources of bedrock and surficial geologic map data are heterogeneous in all respects, including digital vector file formats (e.g., shapefiles, ArcInfo Coverages, and File Geodatabases) and the conventions used to represent and store geometries and their attributes. By simply ingesting geological map data into a common GIS environment, a new synthetic data set with wide utility is created. Going one step further by harmonizing map data into the most basic but common core structure (defined by general field type, not by prescriptive definitions of fields and their contents, see below) requires some effort, but it is also a straightforward task. Currently, Macrostrat's harmonized map database is logically partitioned into four arbitrary map scales (Figure 4) for the purposes of convenience and enhancing the query performance of the system. Despite this scale-based separation at the database-level, each polygon, line, and point ingested into the harmonized data set acquires an internally unique identifier and maintains key-based links back to all original map data.

Lines (e.g., faults, dikes, and fold axes) are not required for a map source, but when they are present, a similar convention is followed. Long-form original descriptions, when applicable, are preserved but a standardized field describing line type (e.g., "normal fault") is always designated or created and then populated. One problem that is unique to lines is the asymmetry that they can contain (e.g., the side of a line that the "teeth" appear on in a thrust fault, which indicate which is the overriding block). There are no widely used protocols for identifying such asymmetries on vector lines and many map sources simply do not contain any direct digital information relevant to line asymmetry. Macrostrat's standardized line structure does allow for the specification of line asymmetry (by convention, the point defining the start of the line is the reference point and asymmetry is defined in a left-right sense relative to that directionality), but most

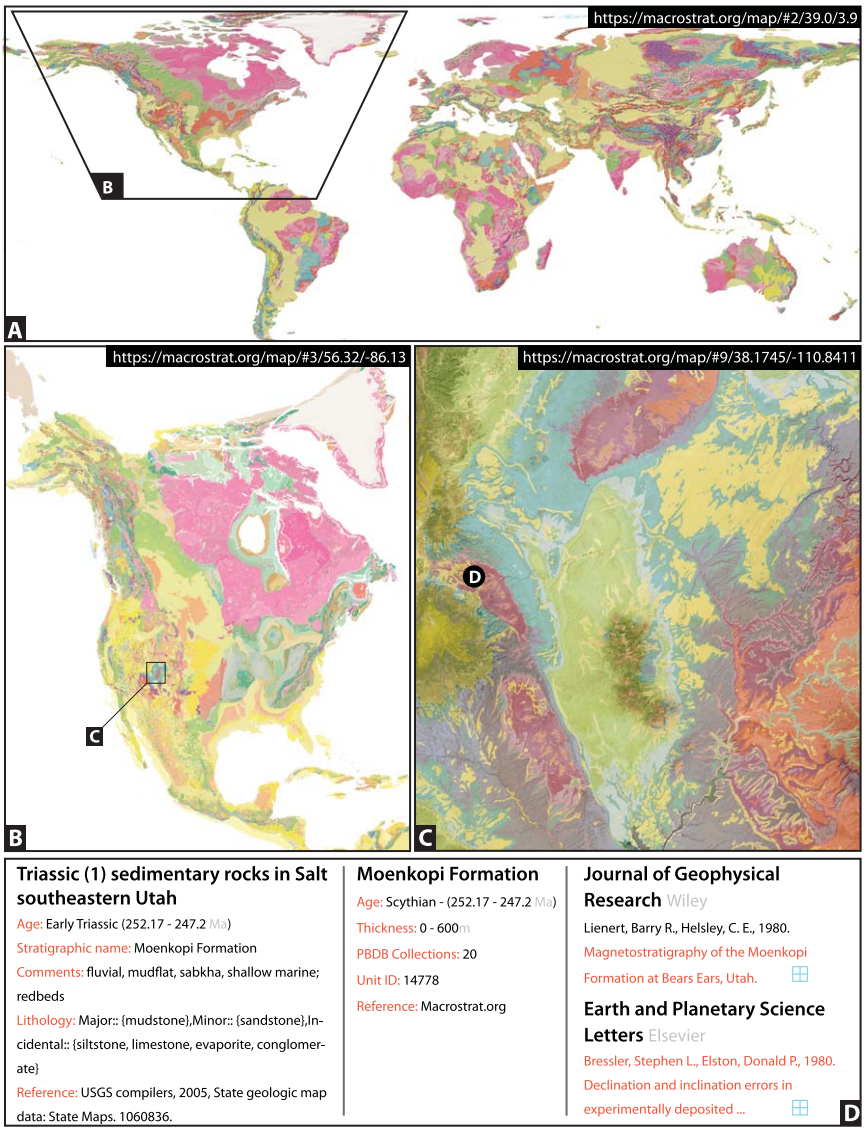

Figure 4. Geological maps at multiple scales and their accessibility from and integration with Macrostrat. A, Generalized geological map of the world (Chorlton, 2007). B, Geologic Map of North America (Garrity & Soller, 2009). C, Geologic map of Utah (Horton et al., 2017; Ludington et al., 2005). D, Result summary obtained by clicking on the map at location of point D. A summary of some of the original map data is shown in left plot; middle plot contains Macrostrat-derived data matched to that map polygon; right plot shows example literature data obtained by using stratigraphic name to identify content in the GeoDeepDive infrastructure.

sources require manual revision, and that process remains incomplete for some maps without substantial loss of information. Points (e.g., foliation or bedding strike-dip, lineation trend-plunge, mineral occurrence) are also optional data. Standardization of point data typically involves assessing/verifying conventions for dip direction (e.g., implicit use of a "right-hand rule") and normalizing descriptions of point types (e.g., "bedding").

After a map's polygons and, optionally, lines and points, have been imported into the standardized database, links between geologic map objects and Macrostrat objects (units, lithostratigraphic names, lithologies, etc.) are established using a combination of spatial and temporal intersection and simple string matching (Figure 4d). The link between map polygons and Macrostrat units is the most complicated step, as it involves: (1) analyzing the strings that are used to name rock units (a step that is informed by the nomenclatural hierarchies), (2) quantitatively assessing the spatial intersections/distances between

Macrostrat units and geological map objects, and (3) testing for overlap in the stated geological ages of each. A geologic map polygon and a Macrostrat unit that overlap in all three attributes (geography, age, and name) constitutes the highest confidence match. Relaxing one or more of these congruences might reduce confidence in the match, but it may still be valid (e.g., a Macrostrat unit and a geologic map polygon may not intersect spatially, but they may be separated by a short distance and share all other attributes, making them highly probable matches). Matches between Macrostrat units and map polygons, and the basis for them, are made algorithmically, but it is also possible to manually remove and add matches.

Although explicit links between geologic map polygons and Macrostrat units can have ambiguity (e.g., due to differences in the way lithostratgraphic units are grouped in a map source versus in Macrostrat), the process is streamlined and the results are repeatable. There also tends to be a large amount of consistency between geologic maps and Macrostrat units because the language that is used to describe rocks in the field is often congruent, at least over the past several decades in many areas of North America. Because spatial expression of rock units is (at least in principle) more objectively defined than estimates of their age or interpretations of their origin, many potential ambiguities that could occur are removed by quantitative tests for spatial and basic descriptive overlap. The end result of matching map units to Macrostrat column units benefits both data sets. Geological map polygons inherit the relevant Macrostrat unit(s) modeled ages (Figure 3), which are often more precise than the geochronological interval that are commonly designated on geological map polygons. Similarly, any other data that have been linked to a Macrostrat unit, such as PBDB fossil occurrences or paleocurrent measurements Brand et al. (2015), can be inherited as attributes of map polygons. Macrostrat units, in turn, benefit by acquiring new information about field properties, including much-needed constraints on their surface expression and physical contact relationships and more complete, first-hand field descriptions of lithology and other attributes.

Macrostrat's geologic map coverage is globally complete at the smallest map scale (Figure 4a), but larger scale coverage is patchy geographically, and it will always be so because that is the nature of the way geological maps are produced. Nevertheless, there are currently some 2.3 million geologic map polygons from more than 200 distinct sources globally integrated in a seamless "Google Maps-like" environment (see http://macrostrat.org/map/sources for a complete listing and spatial index). More than 15,000 Macrostrat units in the regions covered by columns (Figure 2) have been matched to bedrock polygons. The process of adding new geological map data and linking relevant data to Macrostrat is well-defined. Once a map is added to the system, validated, and then transferred to the primary server, all data automatically propagate throughout the entire system (e.g., all new map data automatically show up in the web-based map viewing application accessible at https://macrostrat.org/map, as well as third-party applications).

Expanding and improving the geological map data set is currently limited by the time required to find, download, and import geological maps into Macrostrat's GIS environment. Some geological maps have also not yet been made publicly available in a vector-based format or, if the data are available, they are not public or have licensing terms that prohibit their modification and reuse. The latter is particularly regrettable because geological maps are usually produced with public funds and represent important, basic geological field data that are often underutilized.

2.7. Topographic Data

Bedrock and surficial geology are intimately connected to Earth's surface topography. For this reason, we have integrated NOAA's ETOPO1 (Amante & Eakins, 2009) and the most recent release of NASA/JPL's Space Shuttle Radar Topography Mission (SRTM) data (Farr et al., 2007) into the Macrostrat geological map infrastructure. It is notable that both of these data sets are raster based, rather than vector based, illustrating that Macrostrat (by virtue of its GIS underpinnings) is capable of harnessing any type of geospatial data. Elevation can also, therefore, be readily intersected with other Macrostrat data (e.g., bedrock geologic maps). Although not yet extensively utilized in Macrotrat's public applications, topographic data are accessible in some basic capacities in the mobile application, Rockd, described below, as well as within Macrostrat's web interface.

2.8. GPlates Plate Tectonic Rotations

Paleogeographic context is critical to many questions in historical Earth systems science (Berry & Wilkinson, 1994; Valentine & Moores, 1970; Walker et al., 2002; Zaffos et al., 2017). Geological data, in turn, provides fundamental constraints on paleo-reconstructions (Cao et al., 2017; Wright et al., 2013).

Adapted versions of the GPlates software, and associated rotation models from various authors (Matthews et al., 2016; Merdith et al., 2017; Müller et al., 2016; Seton et al., 2012; Williams et al., 2012; Wright et al., 2013) are used to provide a working plate tectonic rotation model for all Macrostrat and Macrostrat-linked data, such as maps and fossil collections, back to 560 Ma. These rotation models are run on Macrostrat infrastructure using an adapted PyGPlates Python package, which enables application of different reconstruction models and ages to myriad data. This rotation model is currently deployed as a web service and it powers location-aware interactivity in mobile applications (see below).

3. Application Programing Interface (API)

Database design is a critically important component of any data infrastructure. Decisions at this level ultimately impact the efficiency and reliability of data entry, editing, and retrieval. Most of Macrostrat's data, outlined in general terms above, are stored in approximately third normal form relational database structure (see Figure 5 for a simplified schematic). However, modern methods of accessing data that are housed on remote repositories typically do not require any detailed knowledge of database design or software. Specifically, APIs provide a set of tools for building software, and in the context of databases, they provide a specification for how to make remote requests for data using a standard protocol (usually HTTP) and a parameterization that does not depend upon knowledge of underlying database software, schemas, or server-executed code. The remote server's responses to such requests are also formatted using standards that are not specific to any one end use. The general principles governing the deployment of APIs vary, but most modern examples follow a Representation State Transfer (REST; Fielding, 2000) model. Although there are few widely agreed upon implementation details of a REST-ful system, one of the principles is the identification of data resources using Uniform Resource Identifiers (URIs), for example:

https://macrostrat.org/api/v2/defs/strat_names?strat_name=waldron&rank=fm

This URL (Uniform Resource Locator, a type of URI) returns metadata that is specific to the API (version number and data license), along with relevant data, which in this case is basic summary information for all lithostratigraphic names of formation rank that have a name string matching "waldron" (case insensitive).

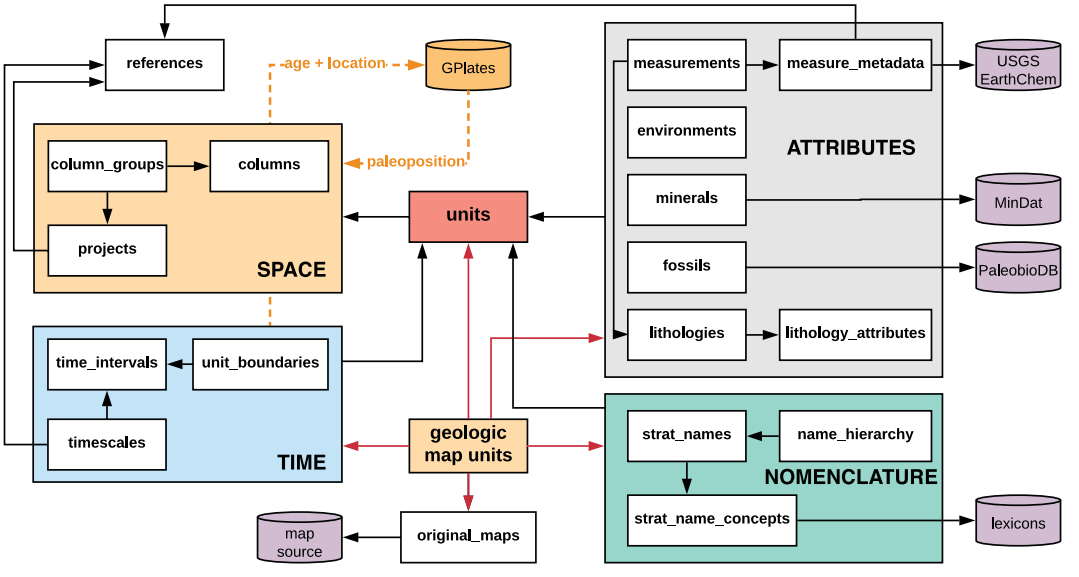

Figure 5. Simplified schematic of core database elements and their relationships in Macrostrat. Columns (as in Figure 1) store spatial data and group one or more units. Purple cylinders represent external database resources. Orange cylinder represents GPlates plate rotation model. Intermediate join tables as well as other internal tables, relationships, and table fields are omitted for clarity. Black arrows identify general table relations stored within the same relational database, red arrows identify table relations across relational databases, dotted orange arrows show flow of information from Macrostrat to-from GPlates rotation model.

Table 1
Select Macrostrat API Routes Available in Version 2

Route	Formats	Description
/columns	json, csv, geojson	Search and summarize columns based on unit propertiesor geographic location
/sections	json, csv	Summarize units by gap-bound packages
/units	json, csv, geojson	Search and summarize units based on their properties
/defs/lithologies	json, csv	Rock types and hierarchies
/defs/lihtology_attributes	json, csv	Modifiers applied to rock types
/defs/environments	json, csv	Depositional environments and hierarchies
/defs/strat_names	json, csv	Lithostratigraphic names and hierarchy
/defs/strat_names_concepts	json, csv	Grouping, attributes and sources for strat_names
/defs/intervals	json, csv	Chronostratigraphic time intervals
/defs/time scales	json, csv	Chronostratigraphic time scales
/defs/measurements	json, csv	Measurements and measurement groups
/defs/minerals	json, csv	Mineral names and chemistries
/geologic_units/map	json, csv, geojson	Geologic map data for lat-lng coordinate or stratigraphic name

Note. Each route described here is preceded by the base URL https://macrostrat.org/api/v2, which also returns this table in expanded, JSON format. Omitting a version in the base URL (i.e., v2) defaults to the latest version of the API. The version number should be included in the URL to ensure that a given API call behaves consistently as the API is updated and modified. For information on parameters accepted by each route and its response, visit the base route (e.g., https://macrostrat.org/api/v2/defs/lithologies).

Lithological nomenclatural concepts, as described above, are a different object and therefore have a different URL, for example:

https://macrostrat.org/api/v2/defs/strat_name_concepts?name=waldron

This returns data for all lithostratigraphic concepts with names matching the string "waldron" (case insensitive) including a unique identifier for each concept and additional information about age, usage, source information, and any available links back to original resources (e.g., USGS Lexicon).

Most responses returned by the Macrostrat API are, by default, formatted in JavaScript Object Notation (JSON), a platform-independent, open standard format. Responses formatted as comma-separated values (CSV) can also be obtained for most routes by adding the parameter "&format = csv" to the URL, though doing so can result in complex fields due to the inability of CSV to readily accommodate nested hierarchies. Routes returning geographic objects (i.e., points, lines, and polygons) can also be formatted as GeoJSON or TopoJSON by supplying an appropriate "&format=" parameter.

Basic documentation for each route in the Macrostrat API is accessible by invoking the base URL (Table 1). For example, general information about the Macrostrat API as a whole and all available routes are returned, in JSON format, by https://macrostrat.org/api. The base URL of each listed route (e.g., https://macrostrat.org/api/columns) returns simple documentation that is specific to the given route, including accepted parameters, available response formats, and brief explanations for the returned fields and their values.

API requests can be generated, made, and processed automatically in any programing environment that is capable of making and receiving HTTP requests. Examples of such environments currently in wide use among geoscientists include R, Python, and Matlab. Figure 6a shows one such example in which the abundance of coal is quantified as a time series in Matlab by requesting the appropriate data via the Macrostrat API and plotting the results using Matlab's built-in plotting functions (with additional customizations). To make this figure, North American units that contain any amount of organic sediment are first requested by properly formatting an API URL:

https://macrostrat.org/api/v2/units?lith_type=organic&project_id=1

The continuous-time modeled ages of the units are used to define gap-bound package structure in each column—in this instance, defined as the number of unique columns occupied by coal units in 1 Myr increments. For convenience, Macrostrat includes an API route that generates such package

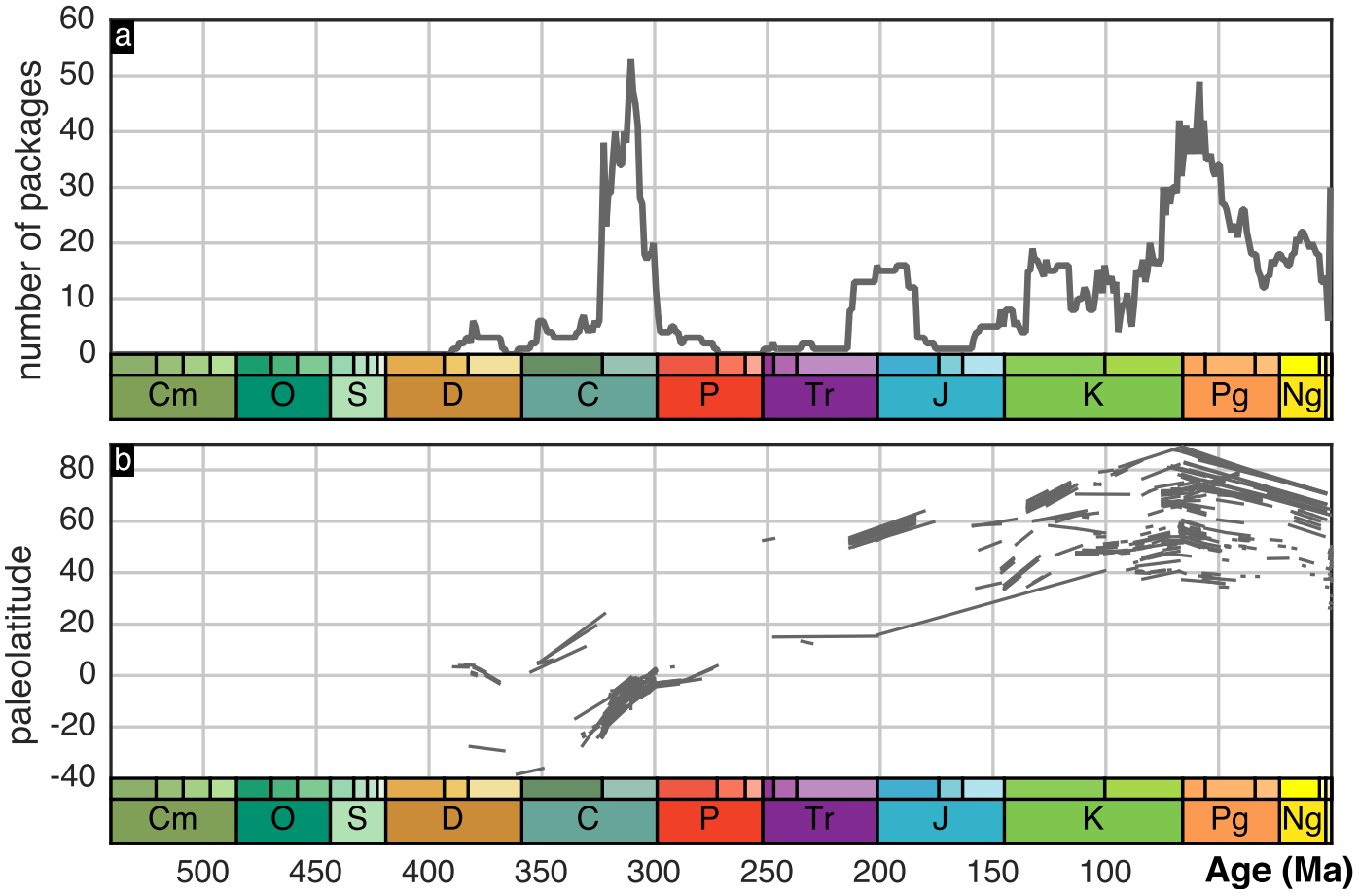

Figure 6. (a) Time series showing the number of coal-bearing and peat-bearing (i.e., "organic") packages (Nelsen et al., 2016). (b) GPlates-modeled paleolatitude versus age for organic sediments. Each line segment shows the latitudinal position over the duration of a sedimentary unit containing organic sediments.

summaries for the specified subset of units (https://macrostrat.org/api/sections?lith_type=organic). These results are pertinent to our understanding of the mechanisms for Paleozoic coal formation (Nelsen et al., 2016).

Owing to Macrostrat's integration with paleogeographic data and models (see above), the latitudinal distribution of coal-bearing and peat-bearing units could also be analyzed as a time series (Figure 6b). In this instance, the long-form API response would be required because paleogeographic rotations are not included in the short-form API response (to allow the most basic and commonly used data to be retrieved with the lowest network overhead possible):

https://macrostrat.org/api/v2/units?lith_type=organic&project_id=1&response=long

The Macrostrat API can also be used to obtain other relevant data, such as international chronostratigraphic period names, abbreviations, ages and their conventional colors. This API call is in fact used to generate the graphical time scales in Figure 6. The Macrostrat API can also be used to supply geospatial data in a GIS software environment (e.g., QGIS).

4. Example Applications

The power of APIs is that they allow data to be accessed using a common protocol but analyzed and displayed in many different ways. Several mobile and web applications that use the Macrostrat API are now publicly available, including the iOS and Android application Flyover Country and the iOS application Mancos, each developed by third parties. Here we briefly describe the Rockd mobile application developed by the Macrostrat group.

Rockd (https://rockd.org) is built using the Ionic framework and it leverages Macrostrat's geologic map data as well as lithostratigraphic nomenclature, lithologies, paleogeographic reconstructions, and more. One of the fundamental questions that Rockd aims to help users answer is, "what rock am I standing on, and where and when did it form?" Finding the answer to such a question previously required either knowledge and direct observation or searching for a scale-appropriate geological map, viewing the map, and then estimating a location on that map relative to landmarks or a GPS device's coordinates. Then, once a geological age and context was acquired from a map or other published source, a user would still typically have to locate and consult other sources for a reconstructed paleogeographic position at the time of the rock's formation.

Macrostrat's data infrastructure allows users to answer the "what am I standing on" question in real time anywhere in the world, with levels of detail that vary regionally but using a platform that continually improves. This same data infrastructure also allows users to obtain their current elevation using digital elevation models (ETOPO1 and SRTM1) instead of the elevation reported by their device's GPS chip, which is prone to large errors in vertical position (GPS-specific and device-specific uncertainty in horizontal position still occurs and will affect elevation estimates; precision of the GPS-supplied latitude-longitude estimate is reported in Rockd). Additionally, by retrieving data from a web API wrapper of GPlate's (Wright et al., 2013) PyGPlates Python package, users are automatically given their paleoposition for any time going back to 750 Ma. Global paleogeographic reconstructions and the user's paleoposition are then shown on paloegeographic reconstructions from C. R. Scotese (Scotese, 2016). Reliance on Macrostrat's API (and Rockd's own API) allows the application to be continually updated without any user intervention (e.g., a new geological map added to Macrostrat will be accessible without requiring installation of a new version of the application).

In addition to providing local geological context, Rockd allows registered users to record their own field observations and to make them public on Macrostrat servers. User observations can leverage existing geological knowledge, such as stratigraphic names, lithologies, and taxonomic names that are known to occur around the observation's location. Providing this type of location-specific information can speed up the data acquisition process and improve the quality of data by reducing the need to type text. Delivering local geological context also, in principle, encourages users to focus their efforts on making observations that might supply new information or that complements or revises existing information, thereby enhancing local geological knowledge and ultimately improving rock unit descriptions. In keeping with REST principles, Rockd photos, observations and locations are assigned URLs that, when made public by a user, can be shared (e.g., https://rockd.org/checkin/1727) and commented on by other registered users, offering a mechanism to label locations with alternative interpretations and encourage a learning dialogue. User-contributed checkins (Rockd's term for a location with one or more observations) can also help streamline field work and the planning of field trips. The ability to create custom, ordered groupings of locations that can be named and identified by a single URL is forthcoming and will serve as the organized "field trip" component of Rockd.

Rockd is a mobile app that draws on Macrostrats API for data exploration and visualization purposes, but the Macrostrat data service can also be used in geoscientific applications. Figure 6 shows output from one such scientific application. In this example, unit data in either JSON or CSV format are first retrieved from the Macrostrat API using the URL above. These data are then parsed and for each time increment of interest (here 0–541 Ma) and the number of units with overlapping ages are then counted by looping over all units in the response. If the fields with the continuous-time age estimate for each unit are used ("b_age" and "t_age"), then this estimate can be a time "interval-free" summarization of rock quantity through geologic time. Other parameters, such as rock area and volume can be summarized in the same capacity, although the latter requires the user to make decisions about how to include proportional lithological abundances and the maximum and minimum thickness estimates that accompany each unit.

5. Example Results and Future Directions

Macrostrat is, first and foremost, designed to generate novel results that can be used to test geological hypotheses about rock preservation and cycling (Peters, 2006; Peters & Husson, 2017) and the drivers of biological (Hannisdal & Peters, 2011; Peters, 2008a, 2008b; Peters et al., 2013; Peters & Gaines, 2012) and

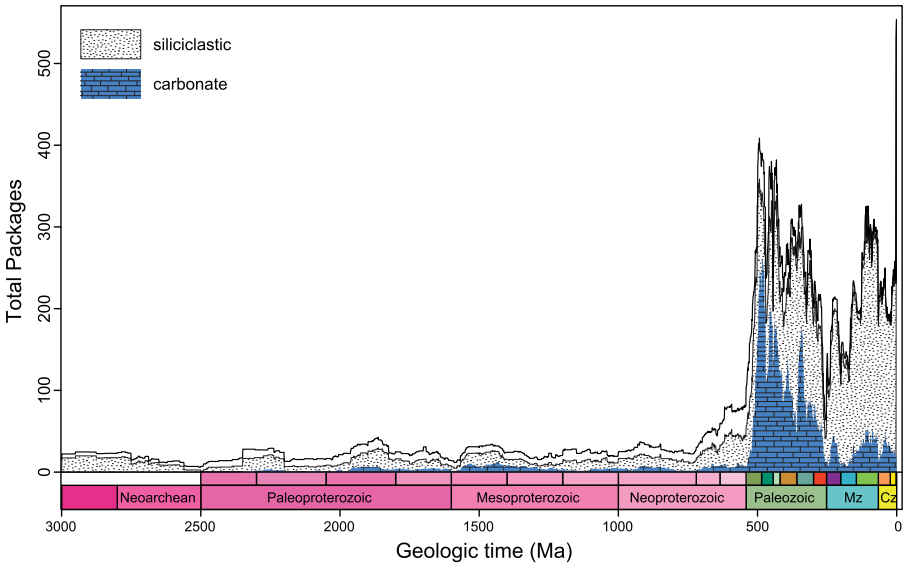

Figure 7. Total number of sedimentary (including metasedimentary) rock packages in North America (Figure 1; a total of 1,013 columns are present in this area). Lithologies are subivided into three groups: carbonates, siliciclastics, and all others (no color or fill). Units with multiple lithologies are weighted according to the proportion of each group (e.g., a package with a unit composed of 50% carbonate and 50% sandstone would contribute 0.5 package units to each lithology type).

biogeochemical (Halevy et al., 2012; Husson & Peters, 2017) evolution. Recent implementation of the preliminary continuous-time age model (Figure 3) has enabled us to conduct substantive quantitative analyses of the Precambrian sedimentary rock record and to compare that record in a meaningful way to the Phanerozoic. Shifting to a time interval-free approach to measuring rock quantity was important because the much longer subdivisions of Precambrian time impart a strong signal when conducting interval-based analyses (Sadler, 1981).

The longest-term history of sediment quantity in the area covered by Macrostrat (Figure 2) is remarkable in several different ways (Figure 7). Most notably, the step-wise increase in sediment quantity across the Precambrian-Cambrian boundary marks what has been called the "Great Unconformity" (Karlstrom & Timmons, 2012; Peters & Gaines, 2012). Determining whether or not this strong signal of increased continent-hosted sediment quantity is North America-specific or a global phenomenon is critical to addressing many fundamental questions about the evolution of Earth and life (Husson & Peters, 2017).

A current major limitation of Macrostrat is the geographically restricted nature of its surface-subsurface data (i.e., columns). Currently, Macrostrat columns cover approximately 15% of the global continental crust, most of which is in North America (Figure 2). Although evidence suggests that there is indeed a global expression of the Great Unconformity (He et al., 2017; Husson & Peters, 2017), testing this hypothesis requires geographic expansion of column coverage. We hope that this objective will be facilitated by engaging geoscientists with regional expertise and by leveraging their in-hand knowledge. Because the units comprising each column in Macrostrat can, at least initially, comprise only the most basic information on lithology, age, and thickness of geological units, many regional geoscientists now have the necessary knowledge and data in-hand to rapidly expand geographic coverage. The accuracy and precision of the general summaries can be improved once the scaffolding that completely describes the upper crust is in place.

To facilitate the task of geographic expansion, some of the basic data required for column entry has already been incorporated into Macrostrat. For example, Australian geology is represented by geological maps at multiple scales and by the entire Australian stratigraphic lexicon. These data, in combination with definitions of lithology, lithology attributes, and chronostratigraphic intervals already in Macrostrat, mean that the process of entering a new column in Australia would require (1) defining the geographic region of interest by designating a bounding geometry and (2) defining a chronostratigraphic succession of units that are linked to lithologies, thicknesses and, optionally, lithostratigraphic names and environments of formation.

In order to obtain the highest-quality initial columns, participation of regional experts is the ideal path forward. Launching a globally comprehensive initiative to harness regional geological expertise and synthesize field experience and knowledge would have many far-reaching, positive impacts, including enabling key hypotheses to be tested and establishing a digitally accessible, comprehensive working model of the age and material properties of rocks in the Earth's upper crust.

Acknowledgments
All Macrostrat data are available via the API at https://macrostrat.org. Grants from National Science Foundation, the American Chemical Society, and the USGS supported initial data compilation and analysis during the period from 2006 through approximately 2010. Major recent data infrastructure development supported by US National Science Foundation EAR-1150082 and EarthCube grant ICER-1440312. GeoDeepDive infrastructure development supported by NSF's EarthCube ICER-1343760. We thank Matthew Kosnik, Fred Ziegler, Bruce Wilkinson, Michael Foote, Alan Carroll, Steve Meyers, Michael McClennen, Steve Holland, and Andrew Zaffos for their input and feedback on macrostratigraphy methodology and on database development. Noel Heim, Deborah Rook, Sharon McMullen, Neal Auchter, and Annaka Clement aided in data entry and editing. Noel Heim and Puneet Kishor provided invaluable contributions to the database and initial efforts to integrate geological maps. We thank Wolfgang Kiessling and an anonymous reviewer for helpful feedback. We also thank the Paleobiology Database contributors; this is Paleobiology Database Official Publication NUM.

References

Amante, C., & Eakins, B. W. (2009). *ETOPO1 1 arc-minute global relief model: Procedures, data sources and analysis*. US Department of Commerce, National Oceanic and Atmospheric Administration, National Environmental Satellite, Data, and Information Service, National Geophysical Data Center, Marine Geology and Geophysics Division Colorado.

Aswasereelert, W., Meyers, S. R., Carroll, A. R., Peters, S. E., Smith, M. E., & Feigl, K. L. (2013). Basin-scale cyclostratigraphy of the Green River Formation, Wyoming. *Geological Society of America Bulletin*, *125*(1–2), 216–228.

Benton, M. J., Donoghue, P. C., Asher, R. J., Friedman, M., Near, T. J., & Vinther, J. (2015). Constraints on the timescale of animal evolutionary history. *Palaeontologia Electronica*, *18*(1), 1–106.

Bergman, N. M., Lenton, T. M., & Watson, A. J. (2004). COPSE: A new model of biogeochemical cycling over Phanerozoic time. *American Journal of Science*, *304*(5), 397–437.

Berner, R. A. (1989). Biogeochemical cycles of carbon and sulfur and their effect on atmospheric oxygen over Phanerozoic time. *Global and Planetary Change*, *1*(1–2), 97–122.

Berner, R. A., & Canfield, D. E. (1989). A new model for atmospheric oxygen over Phanerozoic time. *American Journal of Science*, *289*(4), 333–361.

Berner, R. A., & Kothavala, Z. (2001). GEOCARB III: A revised model of atmospheric CO_2 over Phanerozoic time. *American Journal of Science*, *301*(2), 182–204.

Berry, J. P., & Wilkinson, B. H. (1994). Paleoclimatic and tectonic control on the accumulation of North American cratonic sediment. *Geological Society of America Bulletin*, *106*(7), 855–865.

Brand, L., Wang, M., & Chadwick, A. (2015). Global database of paleocurrent trends through the Phanerozoic and Precambrian. *Scientific Data*, *2*, 150025.

Budyko, M. I., Ronov, A. B., & Yanshin, A. L. (1987). *History of the Earth's atmosphere*. Berlin, Germany: Springer.

Cao, W., Zahirovic, S., Flament, N., Williams, S., Golonka, J., & Müller, R. D. (2017). Improving global paleogeography since the late Paleozoic using paleobiology. *Biogeosciences*, *14*(23), 5425–5439.

Childs, O. E. (1985). Correlation of stratigraphic units of North America—COSUNA. *AAPG Bulletin*, *69*(2), 173–180.

Chorlton, L. B. (2007). Generalized geology of the world: Bedrock domains and major faults in GIS format: A small-scale world geology map with an extended geological attribute database. *Geological Survey of Canada Open File*, *5529*, 48.

Des Marais, D., Strauss, H., Summons, R., & Hayes, J. (1992). Carbon isotope evidence for the stepwise oxidation of the Proterozoic environment. *Nature*, *359*(6396), 605.

Fan, Y., Richard, S., Bristol, R., Peters, S. E., Ingebritsen, S., Moosdorf, E., et al. (2016). Digitalcrust—A 4D data system of material properties for transforming research on crustal fluid flow. In *Crustal permeability* (pp. 6–12). Hoboken, NJ: John Wiley.

Farr, T. G., Rosen, P. A., Caro, E., Crippen, R., Duren, R., Hensley, S., et al. (2007). The shuttle radar topography mission. *Reviews of Geophysics*, *45*, RG2004. https://doi.org/10.1029/2005RG000183

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* (doctoral dissertation).

Fraass, A. J., Kelly, D. C., & Peters, S. E. (2015). Macroevolutionary history of the planktic foraminifera. *Annual Review of Earth and Planetary Sciences*, *43*(1), 139–166.

Garrity, C., & Soller, D. (2009). Database of the Geologic Map of North America; adapted from the map by J. C. Reed, Jr. and others (2005), *Data series 424*. Reston, VA: U.S. Geological Survey.

Gleeson, T., Befus, K. M., Jasechko, S., Luijendijk, E., & Cardenas, M. B. (2016). The global volume and distribution of modern groundwater. *Nature Geoscience*, *9*(2), 161–167.

Gombosi, D. J., & Wilkinson, B. H. (2012). Global rates of geologic cycling: Tectonic diffusion of upper crustal lithosomes. *The Journal of Geology*, *120*(2), 121–133.

Gradstein, F. M., Ogg, G., & Schmitz, M. (2012). *The geologic time scale 2012*. Amsterdam, the Netherlands: Elsevier.

Halevy, I., Peters, S. E., & Fischer, W. W. (2012). Sulfate burial constraints on the Phanerozoic sulfur cycle. *Science*, *337*(6092), 331–334.

Hannisdal, B., & Peters, S. E. (2010). On the relationship between macrostratigraphy and geological processes: Quantitative information capture and sampling robustness. *The Journal of Geology*, *118*(2), 111–130.

Hannisdal, B., & Peters, S. E. (2011). Phanerozoic earth system evolution and marine biodiversity. *Science*, *334*(6059), 1121–1124.

Hay, W. W., & Leslie, M. A. (1990). Could possible changes in global groundwater reservoir cause eustatic sea-level fluctuations. In *Sea-level change* (pp. 161–170). Washington, DC: NRC.

He, T., Zhou, Y., Vermeesch, P., Rittner, M., Miao, L., Zhu, M., et al. (2017). Measuring the 'Great Unconformity' on the North China craton using new detrital zircon age data. *Geological Society, Special Publications*, *448*(1), 145–159.

Horton, J. D., San Juan, C. A., & Stoeser, D. B. (2017). The State Geologic Map Compilation (SGMC) geodatabase of the conterminous United States (ver. 1.1, August 2017). *United States Geological Survey data series*, 1052, 46 pp.

Husson, J. M., & Peters, S. E. (2017). Atmospheric oxygenation driven by unsteady growth of the continental sedimentary reservoir. *Earth and Planetary Science Letters*, *460*, 68–75.

Karlstrom, K. E., & Timmons, J. M. (2012). Many unconformities make one 'Great Unconformity.' *Geological Society of America Special Papers*, *489*, 73–79.

King, P. R., Naish, T. R., Browne, G. H., Field, B. D., & Edbrooke, S. W. (1999). *Cretaceous to recent sedimentation in New Zealand, Folio series* (vol. 1). Lower Hutt, New Zealand: Institute of Geological & Nuclear Sciences Limited.

Ludington, S., Moring, B. C., Miller, R. J., Stone, P. A., Bookstrom, A. A., Bedford, D. R., et al. (2005). Preliminary integrated geologic map databases for the United States—Western states: California, Nevada, Arizona, Washington, Oregon, Idaho, and Utah. *United States Geological Survey Open-File Report*, *1305*.

Matthews, K. J., Maloney, K. T., Zahirovic, S., Williams, S. E., Seton, M., & Müller, R. D. (2016). Global plate boundary evolution and kinematics since the late paleozoic. *Global and Planetary Change*, *146*, 226–250.

Maurrasse, F. (1990). Stratigraphic correlation for the circum—Caribbean region. In G. Dengo & J. E. Case (Eds.), *Decade of North American geology, volume H: The Caribbean Region* (Plate 4). Boulder, CO: Geological Society of America Inc.
Merdith, A. S., Collins, A. S., Williams, S., Pisarevsky, E. S., Foden, J., Archibald, F. D., et al. (2017). A full-plate global reconstruction of the Neoproterozoic. *Gondwana Research*, *50*, 84–134.
Meyers, S. R., Siewert, S. E., Singer, B. S., Sageman, B. B., Condon, D. J., Obradovich, J. D., et al. (2012). Intercalibration of radioisotopic and astrochronologic time scales for the Cenomanian-Turonian boundary interval, Western Interior Basin, USA. *Geology*, *40*(1), 7–10.
Mooney, W. D., Laske, G., & Masters, T. G. (1998). Crust 5.1: A global crustal model at 5 × 5. *Journal of Geophysical Research*, *103*(B1), 727–747.
Müller, R. D., Seton, M., Zahirovic, S., Williams, S. E., Matthews, K. J., Wright, N. M., et al. (2016). Ocean basin evolution and global-scale plate reorganization events since Pangea breakup. *Annual Review of Earth and Planetary Sciences*, *44*(1), 107–138.
Nelsen, M. P., DiMichele, W. A., Peters, S. E., & Boyce, C. K. (2016). Delayed fungal evolution did not cause the Paleozoic peak in coal production. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(9), 2442–2447.
Peters, S. E. (2006). Macrostratigraphy of North America. *The Journal of Geology*, *114*(4), 391–412.
Peters, S. E. (2008a). Macrostratigraphy and its promise for paleobiology. In P. H. Kelley & R. K. Bambach (Eds.), *From evolution to geobiology: Research questions driving paleontology at the start of a new century, Paleontological Society Papers* (Vol. 14, pp. 205–232).
Peters, S. E. (2008b). Environmental determinants of extinction selectivity in the fossil record. *Nature*, *454*(7204), 626–629.
Peters, S. E., & Gaines, R. R. (2012). Formation of the 'Great Unconformity' as a trigger for the Cambrian explosion. *Nature*, *484*(7394), 363–366.
Peters, S. E., & Heim, N. A. (2010). The geological completeness of paleontological sampling in North America. *Paleobiology*, *36*(01), 61–79.
Peters, S. E., & Husson, J. M. (2017). Sediment cycling on continental and oceanic crust. *Geology*, *45*(4), 323–326.
Peters, S. E., Kelly, D. C., & Fraass, A. J. (2013). Oceanographic controls on the diversity and extinction of planktonic foraminifera. *Nature*, *493*(7432), 398–401.
Raup, D. M. (1976). Species diversity in the Phanerozoic: An interpretation. *Paleobiology*, *2*(4), 289–297.
Raymond, O., Liu, S., Gallagher, R., Highet, L., & Zhang, W. (2012). *Surface geology of Australia, 1:1 000 000 scale* (2012 ed.) [digital dataset] (technical report). Canberra, Australia: Geoscience Australia, Commonwealth of Australia.
Ronov, A. B. (1982). The Earth's sedimentary shell (quantitative patterns of its structure, compositions, and evolution). *International Geology Review*, *24*(11), 1313–1363.
Ronov, A. B. (1994). Phanerozoic transgressions and regressions on the continents; a quantitative approach based on areas flooded by the sea and areas of marine and continental deposition. *American Journal of Science*, *294*(7), 777–801.
Ronov, A. B., Khain, V. E., Balukhovsky, A. N., & Seslavinsky, K. B. (1980). Quantitative analysis of Phanerozoic sedimentation. *Sedimentary Geology*, *25*(4), 311–325.
Ronov, A. B., & Khain, V. Y. (1954). Deonvian lithologic associations of the world. *Soviet Geology*, *41*, 47–76.
Ronov, A. B., & Migdisov, A. A. (1971). Geochemical history of the crystalline basement and the sedimentary cover of the Russian and North American platforms. *Sedimentology*, *16*(3–4), 137–185.
Ronov, A. B., & Yaroshevsky, A. A. (1969). Chemical composition of the Earth's crust. In H. J. Pembroke (Ed.), *The earth's crust and upper mantle, Geophysical monograph series* (Vol. 13, pp. 37–57). Washington, DC: American Geophysical Union.
Sadler, P. M. (1981). Sediment a ccumulation rates and the completeness of stratigraphic sections. *The Journal of Geology*, *89*(5), 569–584.
Schrag, D. P., Higgins, J. A., Macdonald, F. A., & Johnston, D. T. (2013). Authigenic carbonate and the history of the global carbon cycle. *Science*, *339*(6119), 540–543.
Scotese, C. (2016). *Paleomap PaleoAtlas for GPlates and the PaleoData plotter program*. Retrieved from http://www.earthbyte.org/paleomap-paleoatlas-for-gplates/
Seton, M., Müller, R., Zahirovic, S., Gaina, C., Torsvik, T., Shephard, G., et al. (2012). Global continental and ocean basin reconstructions since 200 Ma. *Earth-Science Reviews*, *113*(3–4), 212–270.
Smith, A. B., & McGowan, A. J. (2007). The shape of the Phanerozoic marine palaeodiversity curve: How much can be predicted from the sedimentary rock record of Western Europe? *Palaeontology*, *50*(4), 765–774.
Stott, D. (1991). Geotectonic correlation chart, Northwest Territories and Yukon, Sedimentary cover of the craton in Canada. In *Geology of North America, D-1, sheet 1*. Geological Society of America.
Trettin, H. (1991). Geotectonic correlation chart 3. In *Geology of Canada*. Geological Survey of Canada.
USGS. (2016). *National geologic map database*.
Valentine, J., & Moores, E. (1970). Plate-tectonic regulation of faunal diversity and sea level: A model. *Nature*, *228*(5272), 657–659.
Walker, L. J., Wilkinson, B. H., & Ivany, L. C. (2002). Continental drift and phanerozoic carbonate accumulation in shallow-shelf and deep-marine settings. *The Journal of Geology*, *110*(1), 75–87.
Wall, P. D., Ivany, L. C., & Wilkinson, B. H. (2009). Revisiting Raup: Exploring the influence of outcrop area on diversity in light of modern sample-standardization techniques. *Paleobiology*, *35*(1), 146–167.
Wilkinson, B. H., McElroy, B. J., Kesler, S. E., Peters, S. E., & Rothman, E. D. (2009). Global geologic maps are tectonic speedometers—Rates of rock cycling from area-age frequencies. *Geological Society of America Bulletin*, *121*(5–6), 760–779.
Wilkinson, B. H., & Walker, J. C. (1989). Phanerozoic cycling of sedimentary carbonate. *American Journal of Science*, *289*(4), 525–548.
Williams, S. E., Müller, R. D., Landgrebe, T. C., & Whittaker, J. M. (2012). An open-source software environment for visualizing and refining plate tectonic reconstructions using high-resolution geological and geophysical data sets. *GSA Today*, *22*(4/5), 4–9.
Wold, C. N., & Hay, W. W. (1990). Estimating ancient sediment fluxes. *American Journal of Science*, *290*(9), 1069–1089.
Wright, N., Zahirovic, S., Müller, R. D., & Seton, M. (2013). Towards community-driven paleogeographic reconstructions: Integrating open-access paleogeographic and paleobiology data with plate tectonics. *Biogeosciences*, *10*(3), 1529–1541. https://doi.org/10.5194/bg-10-1529-2013
Zaffos, A., Finnegan, S., & Peters, S. E. (2017). Plate tectonic regulation of global marine animal diversity. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(22), 5653–5658.