2020年2月，第26卷，第1期，027-043页
February 2020，Vol. 26，No.1， pp. 027-043

高 校 地 质 学 报
Geological Journal of China Universities

DOI: 10.16108/j.issn1006-7493.2019102
引用格式：蒋璟鑫，李超，胡修棉. 2020. 沉积学数据库建设与沉积大数据科学研究进展：以Macrostrat数据库为例[J]. 高校地质学报，26（1）：027-043

沉积学数据库建设与沉积大数据科学研究进展：以Macrostrat数据库为例

蒋璟鑫，李 超，胡修棉*

内生金属矿床成矿机制研究国家重点实验室，南京大学 地球科学与工程学院， 南京 210023

摘要：沉积岩（物）是构成地球表层的主要岩石类型。自地质学诞生以来，地质学家已经积累了海量的沉积学相关研究数据，国内外也相继涌现出Macrostrat等以整合沉积学、地层学相关数据为主的优秀数据库。随着沉积学、地层学、古生物学、地球化学、地质年代学、地球观测等学科数据的快速增长，数据整合分析技术的重大突破，从全球视野研究深时沉积过程变为了可能。文章介绍了国际沉积相关数据库的总体建设情况，并深度剖析美国Macrostrat数据库的结构及其创新工作模式，旨在为深时数字地球（Deep-Time Digital Earth, DDE）计划建设多学科、多尺度、多层次、共享开源的大数据库提供借鉴和参考；在此基础上，剖析了若干应用大数据思维开展的重要科研实例。

关键词：Macrostrat；大数据；数据库；沉积物演化；沉积学
中图分类号：P588.2; P628+.4 **文献标识码：** **文章编号：**1006-7493（2020）01-027-17

Advances on Sedimentary Database Building and Related Research: Macrostrat As an Example

JIANG Jingxin，LI Chao，HU Xiumian*

State Key Laboratory of Mineral Deposit Research, School of Earth Sciences and Engineering, Nanjing University, Nanjing 210023

Abstract: Sedimentary rocks are the main rock type that constitutes the Earth's surface. During centuries a large amount of sedimentological data have been accumulated and in the meanwhile comprehensive sedimentological databases, such as Macrostrat, have established. With the rapid growth of data in all aspects of geology including sedimentology, as well as great breakthroughs in data integration and analysis technology, it is possible to employ big-data analysis methods to explore the deep-time sedimentary process from a global perspective. The currentpaper introduces the main sedimentological databases, and analyzes their structure in detail. The innovative working mode of Macrostrat database is deciphered aiming to provide valuable experience for the sedimentological database in the Deep-time Digital Earth (DDE) Big Science Program. The database will be multi-disciplinary, multi-scaled, multi-leveled and opensource. Several study cases of employing big data analysis to solve scientific questions are also introduced here.

Key words: Macrostrat; big data; database; evolution of sediments; sedimentology

Corresponding author: HU Xiumian, Professor; E-mail: huxm@nju.edu.cn

随着数据存储、运算、分析技术的进步，人类具备了处理海量数据、并从中提取信息的能力，新的科研范式——数据密集型科学研究应运而生。它正在潜移默化地影响着人类生活，改变

收稿日期：2019-11-08；**修回日期：**2019-11-25
基金项目：国家杰出青年基金（41525007）资助
作者简介：蒋璟鑫，男，1995年生，博士研究生，主要从事沉积古环境研究；E-mail: jjxcug24@163.com
***通讯作者：**胡修棉，男，1974年生，教授，主要从事沉积学研究；E-mail: huxm@nju.edu.cn

人类认识和科学研究世界的思维方式（姜浩端，2013；张维明和唐九阳，2015；翟明国等，2018）。地质学的研究突破依赖于对区域或全球各类地质数据的综合分析，是典型的数据密集型科学。在大数据时代，地质学正面临着前所未有的机遇与挑战，地球科学家亟需改变传统的思维方式，从因果关系为核心的逻辑思维方式转变为以关联关系为核心的大数据思维方式（周永章等，2016；陈建平等，2017）。

沉积岩（物）占据了地球表面约70%的面积，是地球表层的重要组成部分。沉积物质作为岩石圈的一部分，其演化受多种地球系统过程控制（生物、气候、构造等），从而忠实地记录了地球表层圈层的演化过程。地球表层沉积物质的总量、类型、通量、时空分布等直接反映了岩石圈、生物圈、水圈、大气圈动态演化的过程，是探讨大尺度时空模式下构造、气候和生物演化的重要参数和基本条件。在20世纪80年代，由全球沉积学家共同发起全球沉积地质计划（Global Sedimentary Geology Program, GSGP[①]），以响应板块学说、古海洋学、古气候学以及沉积地质学等的快速发展，旨在为开展全球尺度的沉积地质研究提供新的方向、机会和动力。基于GSGP，沉积学家提出了三大关键性的研究主题：（1）全球性韵律和事件；（2）全球性演化的沉积学记录；（3）全球性的沉积岩相分析，并将“白垩纪地质记录与全球地质作用、资源、韵律和事件”作为第一个试点项目（陈友明，1987；刘宝珺，1988；叶德燎，1988；Ginsburg, 1986）。这些重大科学问题的提出成为推动沉积学发展的主动力。随着近几十年沉积学、地层学、古生物学、沉积地球化学、地质年代学、地球观测等学科的进一步发展，地质学家积累了海量的沉积学相关的数据。如何高效地整合各类数据，并从中挖掘这些数据中的价值，已经成为沉积学家急需解决的新时代课题。

1 国际沉积相关数据库建设情况

目前，国际上已涌现出一大批优秀的沉积学相关数据库，如Macrostrat、GeoChron、SedDB、Ava Clastics，以及各种以文献形式发表的数据集，如世界古水流数据集（Brand et al., 2015）、世界洋底沉积物数据集（Dutkiewicz et al., 2015）、世界气候敏感性沉积物数据集（Boucot et al., 2013; Cao et al., 2018）、陆相冲积相泥质岩数据集（McMahon et al., 2018）。这些数据库（集）尝试应用大数据思维，从全球视野理解深时沉积物质的演化和循环过程。下面进行详细介绍。

1.1 俄罗斯Alexander Ronov数据库

在20世纪50年代，俄罗斯Alexander Ronov团队开始对地壳岩石的年龄、岩性和体积进行时空综合数据的人工编译工作。他们主要借助于小比例尺（1: 2500万）的地质图及钻井资料，通过相关参数提取和换算，得到岩石总体积、海洋覆盖面积、平均沉降速率、主要岩性组合丰度等数据并编制了显生宙整个过程中这些参数的变化图（Ronov et al.，1969，1980）。该数据库的数据收集过程长达十余年，建设目的是用定量化的方式来探讨岩石、古地理和构造之间的关系和规律，在其建设初期取得了较多的重要研究成果。由于数据获取的局限性，以及严重依赖科学家或团队的个体贡献，Alexander Ronov数据库早已停止发展。

1.2 美国GeoChron和SedDB数据库

GeoChron和SedDB是隶属于EarthChem（Geochemical Databases for the Earth，www.earthchem.org）的与沉积学相关的数据库。EarthChem是一个社区驱动、旨在保存、发现、访问和可视化最广泛和最丰富的地球化学数据的信息网络平台和数据库门户，由美国科学基金委（National science foundation, NSF）资助。

GeoChron（http://www.geochron.org）收集全球沉积岩碎屑矿物年代学数据，以碎屑锆石年龄数据为主；同时捕获其元数据，以允许将来重新计算，并与其它类型的数据集成。该数据库基于网页端口，由哥伦比亚大学进行管理。主要的数据来源有：从已发表文献人工录入、全球科学家的合作贡献以及定年实验仪器的联网自动上传。目前该数据库共收录全球范围内1630个年代学样品，并进行不定时更新（数据来自：http://www.geochron.org）。整体上数据覆盖极不均匀，中国地

① The global sedimentary geology program: report of an international workshop, Fisher Island, Florida, August, 1986.

区仅有约50个样品（数据由本文作者在GeoChron官网统计得到）。用户可以在网页界面根据岩石类型、矿物类型、定年实验方法、地区等参数进行数据筛选，并以HTML、XLS和XML等格式获取数据集。

SedDB（http://www.earthchem.org/seddb）是一个可检索的、以海洋和陆地沉积物地球化学数据为主的关系型数据库，主要根据已发表的文献数据汇编而成。该数据库由美国Lamont-Doherty地球天文台、俄勒冈州立大学、波士顿大学和博伊西州立大学联合开发，由Lamont-Doherty地球天文台负责运营和维护。SedDB汇编了大量地球表层沉积物质的地球化学数据，用于沉积学、地球化学、岩石学、海洋学和古气候研究，同时用于学科教育领域。与GeoChron类似，SedDB也归档了大量的元数据，以便于后期的数据整合、重新计算和分析。截至2013年，该数据库收录了近10400个沉积岩样品的近75万个独立分析数据（数据统计来自：https://en.wikipedia.org/wiki/SedDB），用户可以在web端口根据经纬度、地理位置、样品类型等参数进行数据检索。该数据库2014年以来已停止更新。

1.3　英国Ava Clastics数据库

Ava Clastics（https://www.pds.group/ava-clastics）是一个世界级的沉积学模拟商用数据库，由英国PDS（Petrotechnical Data Systems）集团和利兹大学地球与环境学院合作创立，主要由利兹大学管理。主要收录古代和现代河流、浅海和深海序列的研究实例，作为储层的类似物，并将其数字化（转化为石油行业软件的岩相代码），为能源行业提供服务。根据所收录的数据和应用目的，分为三个子数据库：

（1）FAKTS（Fluvial Architecture Knowledge Transfer System），是利兹大学河流研究小组(FRG，Fluvial Research Group)为主导的、主要存储河流沉积露头数据的关系型数据库，目的是详细描述河流相储层特征并对其中储藏的烃源岩进行预测。目前收录270个河流研究实例，共50544个河流相单元数据（数据来自：https://www.pds.group/ava-clastics/ Databases# FAKTS）；

（2）SMAKS（Shallow Marine Architecture Knowledge System），是利兹大学浅海研究小组（SMRG，Shallow Marine Research Group）为主导的、主要存储浅海沉积露头数据的关系型数据库，目的是数字化浅海沉积体系的所有基本特征并对浅海油气开发和勘探提供模型。目前收录130个研究实例，共14633个浅海相单元数据（数据来自：https://www.pds.group/ava-clastics/ Databases# SMAKS）；

（3）DMAKS（Deep Marine Architecture Knowledge System），主要存储来自古代露头数据和现代深水碎屑岩系统观测数据的关系型数据库，目的是为深水碎屑岩储层的特征识别提供新的定量模型。目前收录66个深海盆地体系研究实例，共9688个深海相单元数据（数据来自：https://www.pds.group/ava-clastics/ Databases# DMAKS）。

除上述数据库外，世界范围内还有很多与沉积学相关的数据库（集）（表1），如以沉积地化数据为主的GSSID（The global sedimentary sulfur isotope database），以露头数据和模拟为主的SAND（Sedimentary ANalogs Database），以及隶属于各个国家的地质调查相关机构的数据库，如英国地质调查局（British Geological Survey, BGS），拥有400多个数据集，如物理数据集（钻孔岩心、岩石、矿物）、文字记录、档案；中国地质调查局（China Geological Survey）自主开发的地质云（Geocloud）涵盖了大量地质图，包括大量地层、沉积相关的数据。

综上，在大数据潮流到来之际，沉积学领域已经涌现了大量优秀的数据库，这些数据库主要关注某一类或某几类数据，依靠人工数字化团队对文献中的数据进行结构化，是利用大数据思维模式拟解决区域、小规模和短时间尺度特定沉积学问题的有效尝试，但是在面临全球、大规模和长时间尺度综合性的科学问题时，这些数据库仍然有很多的局限和不足之处：（1）规模小，数据形式单一，建设和运营多依赖于个人科学家或单个科研团队；（2）发展前景有限，运行状态完全依赖于资助项目的情况，一旦资助结束，数据库即更新停滞；（3）数据覆盖不均匀，数据收集过程受到科学家自身的研究兴趣和主动性的影响；（4）时空分辨率低，无法反映真实的信息；（5）很多文献和数据库资源不开源，难以二次引用和进一步整合。因此，在当前数据更充足、技术更

表1 国际主要沉积学相关数据库（集）

Table 1 Table of major sedimentological database or dataset

数据库名称	网址	开发者运营者	数据类型	数据库建设目的
Macrostrat	https://macrostrat.org	威斯康辛大学 Shanan E. Peters 团队	北美地区地层、岩性、古生物数据	从盆地和大陆尺度对整个地表和地下沉积岩、火成岩和变质岩的组合进行定量的空间和地质年代学分析
Alexander Ronov′s Database	线下静态数据库	Alexander Ronov 团队	由地质图或钻井获得的全球各类沉积物总量和分布的数据	用定量化的方式来探讨岩石、古地理和构造之间的关系和规律
SedDB |EarthChem	http: //www.earthchem . org/seddb	Lamont-Doherty 地球天文台	沉积岩石地球化学数据	收集汇编海洋和大陆沉积物的地球化学数据，用于沉积学、地球化学、岩石学、海洋学和古气候研究，并用于教育目的
GeoChron |EarthChem	http: //www.geochron.org	哥伦比亚大学	全球沉积岩碎屑矿物年龄数据	服务于EathChem和Eathtime，记录地质年代，同时捕获元数据以满足重新计算以及基他数据进行集成
Ava-clastics	http: //www.pds.group/ava-clastics	利兹大学	收录现代、古代的河流、浅海、深海沉积序列实例数据	对不同环境的沉积实例进行分析转换，为能源行业提供服务
LASED	http: //coastal.er.usgs . gov/lased	USGS (美国地质调查局)	路易斯安那州沉积岩和沉积环境数据	提供基于多种底图的地质数据共享平台
SAND	http://www. sedimentaryanalogsdata base.com	科罗拉多矿业大学	沉积岩露头数据和沉积储层模拟数据	通过构建沉积岩系统体系结构、开发和响应变化的预测模型，促进对全球大陆边缘演化过程的科学理解
MARS	http://dbforms ga. gov.au/pls/www/np m.mars.search	澳大利亚地球科学中心	收录澳大利亚海域的现代海相沉积物数据	为沉积动力学、沉积物定量分析、沉积地球化学研究提供数据基础
The global sedimentary sulfur isotope database	http://ww.cet.edu.au/research-projects/special-projects/gssid-global-sedimentary-sulfur-isotope-database	西澳大学 Selvaraja V 团队	收录全球含硫沉积物的年龄和硫同位素相关数据	为科学界提供一个完整和更新的全球数据库，描述沉积岩的多种硫特征随时间的变化
The global paleocurrent database	Doi: 10.1038/sdata 2015.25 (2015).	罗马琳达大学	收集已发表文献中各大陆前寒武和显生宙古水流数据	为盆地分析、烃源岩研究、板块重建和检验全球性构造事件的时间等提供数据基础
Seafloor sediments in the world′s ocean	Doi:10.1130/G36883.1	悉尼大学 Dutkiewicz A 团队	收录大洋钻探原始航次报告中的沉积物数据	了解全球海洋沉积物分布规律，对气候变化及其对海洋环境的影响进行重建和预测
Alluvial mudrock dataset	Doi: 10.1126/science. aan4660	剑桥大学 McMahonW J 团队	收录石炭纪之前全球冲积相泥质岩数据	研究太古代-石炭纪冲积相泥质岩的演化及其控制因素
Climate-sensitive lithologies dataset	Doi: org/10.2110/sepmcsp.11	俄勒冈州立大学 Boucot A J 团队	收录地质历史时期全球气候敏感性沉积岩数据	总结全球古气候带的特征，结合古地理位置重建气候敏感性沉积岩的古纬度，为研究气候分带和气候变化提供依据

先进的条件下，有必要建立更高精度、更全面的地学数据库，更高效地收集和挖掘沉积地质领域的“暗数据”和长尾数据，进一步探索和理解深时地质历史的演化过程和机制。

2 Macrostrat数据库剖析

Macrostrat是一个综合多学科、多尺度、多层次方法的数据共享平台，侧重于定量总结岩石记录时空分布格局，为科学家研究全球沉积岩记录形成和破坏、大规模古生物演化等问题提供了可能（Peters and Husson, 2018），是现阶段沉积大数据库建设的一个范例。这里详细介绍其数据库的结构、创新的工作模式以及相关的研究实例。

2.1 数据库结构

Macrostrat是以沉积学为主的地质数据库，由美国威斯康辛大学Shanan E. Peters团队创立，于2005年正式启动，由NSF资助。是基于MariaDB[①]和PostGIS-enabled PostgreSQL[②]环境开发的关系型

① MariaDB：一种数据库管理系统，由社区开发，与MySQL（目前最常见的开源关系型数据库系统）高度兼容.

② PostgreSQL：一种开源的对象—关系数据库管理系统；PostGIS是PostgreSQL的一个扩展，提供空间对象、空间索引、空间操作函数和空间操作符等空间信息服务功能（https://zh.wikipedia.org/）.

地理空间数据库和辅助性的网络基础设施，可以通过网页进行访问（https://macrostrat.org）。

Macrostrat 目前主要涵盖北美、加勒比、新西兰地区及 IODP 部分研究区的地层数据、PBDB（Paleobiology Database）的化石数据、USGS（United States Geological Survey）的地球化学数据、Mindat 的矿物数据以及涵盖全球范围的地质图数据。Macrostrat 致力于应用这些新的数据来开展研究。

2.2　空间信息

（1）地层柱（Column），是 Macrostrat 的主要空间数据对象，是代表某个区域整体地质概况的地层综合体，最早由美国科学家在编制北美地层对比表（Correlation of Stratigraphic Units of North America，COSUNA）时提出。在 COSUNA 提供的地层对比表中，每个 Column 本质上是一个复合地层柱，代表了整个盆地的综合地质信息。由于不同区域的构造格架不同，因此人为地决定地层柱的分布密度，在构造程度复杂的区域（如大陆边缘）进行加密，以保证获取最有代表性的地质信息（图 1d）。

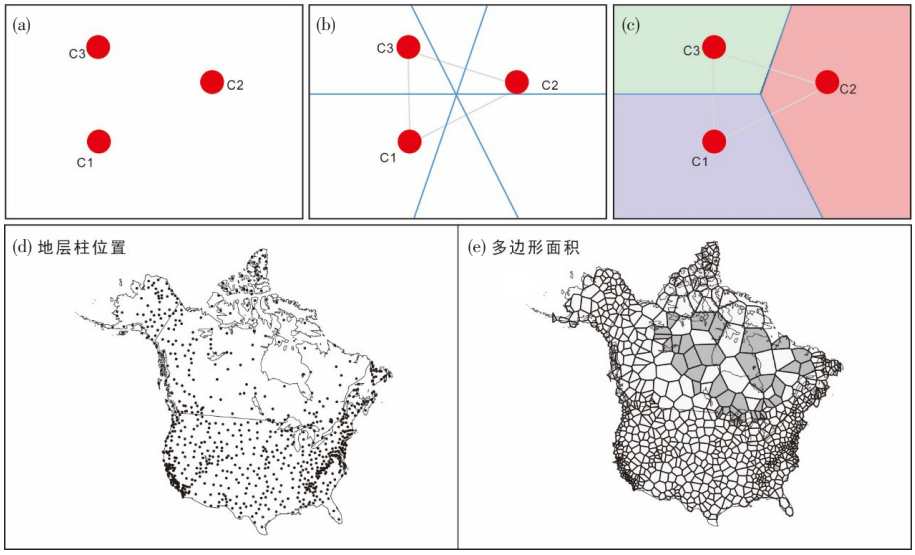
（2）多边形（Polygon），是地层柱映射的地理分区。多边形提出的目的是定量分析整个北美区域的岩石地层信息。Macrostrat 以地层柱为区域岩石地层信息的控制点，按照 Delaunary 三角划分原理（图 1 a–c）为每个控制点分配控制范围，该方法保证了每个多边形内的任意一点与其控制点的距离，都小于与其他控制点的距离，并默认该范围内的地层信息与地层柱一致。该过程是在 R 语言[①]环境下完成的，同时允许对多边形进行人为编辑，以保证多边形的边界与有地质意义的特征边界保持一致，如大的不整合面、断层面等。由于多边形的大小取决于地层柱的密度，因此其大小并不一致（图 1e）。

（3）单元（Units），是组成地层柱的基本元素，也是 Macrostrat 数据库的核心要素，在数据录入时被识别为与其他相邻单元在古生物、岩性和/或年代上不同的岩体或沉积物。在 Macrostrat 中，每个单元具备地层名称、测量数据（如厚度）、沉积环境、矿物、化石、组成单元的岩性（一种或多种）等信息。所有单元属性信息均以表格形式进行存储（图 2）。

2.3　时间信息

2.3.1　地质年代信息

Macrostrat 储存了多种相互关联、在相对和绝对意义上与数值年龄相关的地层划分方案（如年代地层、生物地层、岩石地层等）。其中年代地层单元具有数值年龄，主要参考由国际地层学委员会发布的最新数据（www.stratigraphy.org）；对于没



(a)–(c) Delaunary 三角划分示意图：(a) 地层柱控制点(C1、C2、C3)；(b) 控制点连线（灰色）及垂直平分线（蓝色）；(c) 控制点的控制范围（垂直平分线相交的多边形彩色区域）；(d)–(e) 北美大陆的多边形划分（据 Meyers et al., 2011 改）：(d) 北美地区地层柱的分布位置；(e) 每个地层柱代表的区域

图 1　多边形面积划分原理

Fig. 1　Schematic of polygon areas' division

①R 语言：一种自由软件编程语言与操作环境，主要用于统计分析、绘图、数据挖掘（https://zh.wikipedia.org/）.

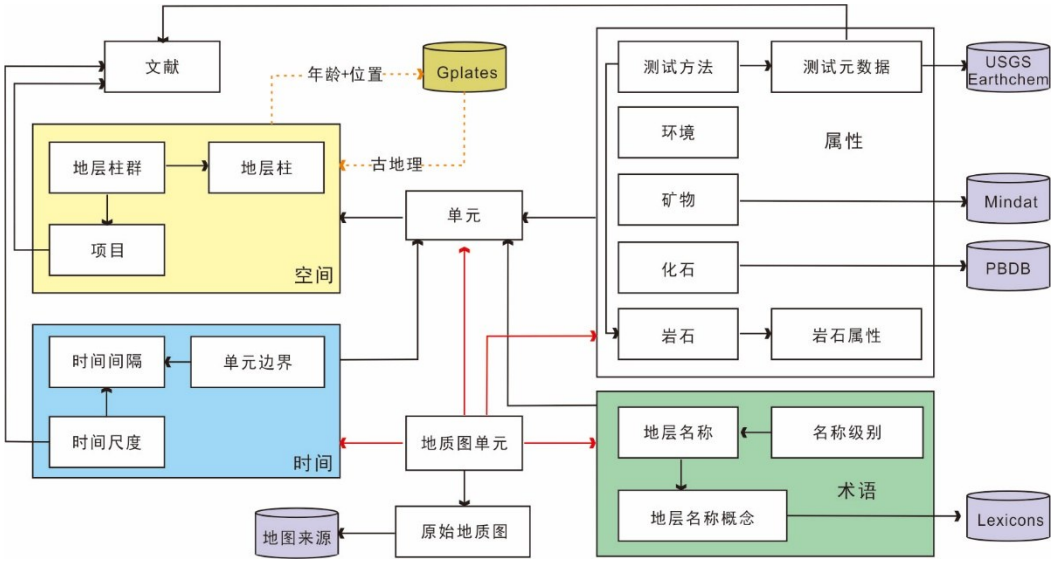


图2 Macrostrat数据库核心元素及其关系示意图（据Peters et al., 2018）

Fig. 2 Simplified schematic of core database elements and their relationships in Macrostrat (from Peters et al., 2018)

有数值年龄限制的地层单元，Macrostrat以相邻地层单元的数值年龄为标尺，按照间隔进行内插标定，系统不直接赋予数值年龄，但其在时间序列上的位置是确定的。这种管理地层划分方案和地质年代信息的方法更加简洁、透明，并具有数据管理优势。

2.3.2 连续年龄模型

传统的地层划分普遍采取“箱式”年龄模型（图3a），即地层单元没有精确数值年龄的限制，而是默认遍历整个地质年代间隔，如图3a中的A单元被限定在整个泥盆系艾菲尔阶，F单元被限定在吉维特阶—弗拉阶。而真实情况是，地层单元的持续时间往往比它们可以相互关联的地质年代间隔要短，因此利用箱式年龄模型进行定量化必定产生较大的误差。

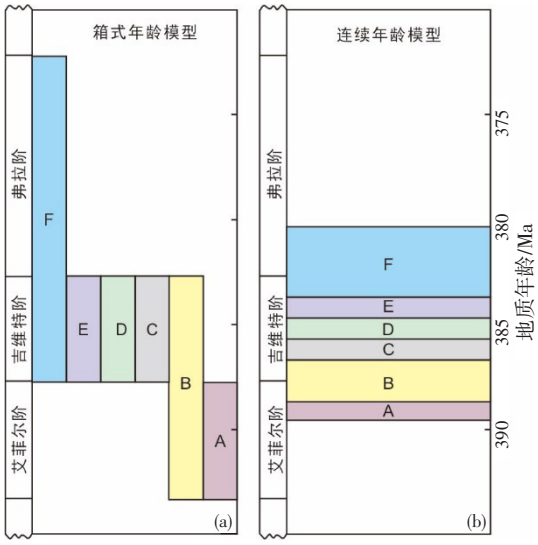


图3 “箱式”年龄模型(a)与连续年龄模型(b)（据Peters et al., 2018）

Fig. 3 (a) “binned” versus (b) continuous age model (from Peters et al., 2018)

为了进行精确的地层量化，Macrostrat提出了地层的连续年龄模型（图3b），（1）在时间轴上，根据古生物谱系、接触关系等时代判断指标，叠加地层单元A–F；（2）选择顶、底具有数值年龄的地层段，对其内部的岩石分布时间进行调整。如已知单元A的底部为389 Ma，并非遍历艾菲尔阶，单元F的顶部为380 Ma，也并非遍历整个弗拉阶，则将A–F限定在389~380 Ma之间；对于无精确年龄限定的BCDE单元，将进行内插标定数值年龄。Macrostrat建立这一模型的目的是进行时间轴上的岩石量化，因此不强调各单元之间的物理接触关系，而强调单元之间的时间连续性。这种沿时间轴以一定时间间隔获取单元数量的量化方式，极大的推动了岩石通量随时间演化的研究。

2.4 岩石地层名称和级别

Macrostrat通过三种方式来管理岩石地层名称：（1）标识相同地质实体的地层名称，如“Dakata砂岩”、“Dakata组”和“Dakata砾岩”，会被分别储存，但指示相同的岩石单元，同时这些名称会与附加信息建立关联，包括地质年龄、地理区

域、参考文献等；（2）对岩石地层名称建立基于从属关系的层级体系，如“Dakata组”是三个“段”级别的更高一级名称，这样便于访问者以任何名称作为关键词访问数据库时，可以获得所有相关的地层数据；（3）通过url来链接相关岩石地层名称术语的原始数据页。

Macrostrat术语管理方式，不仅可以满足岩石地层名称的高效存储，同时由于其岩石地层名称体系的动态性和关联性，数据库能够及时发现潜在的歧义和错误术语并进行改善和补充。

2.5　地质图

Macrostrat嵌入和链接了4种比例尺的全球地质图，目前已涵盖超过200张地质图，超过15000个Macrostrat单元。Macrostrat的地质图数据库存储三种信息：（1）基于矢量的原始地图对象（多边形、直线、点）及其属性，并将其转换为PostGIS环境；（2）所有进行标准化的地图，包括所有地质图对象共有的元素；（3）存储地质图对象和Macrostrat实体的表格。Macrostrat地质图数据的核心是建立地质图多边形与单元之间的联系，同时任何其他与Macrostrat单元相关联的数据，如PBDB化石数据、古水流测量数据等都可以作为地图多边形的属性进行继承，其最终目的是将地质图所包含的资料和信息用于现场地质考察、数据综合分析等。

2.6　地形数据

Macrostrat提供美国国家海洋和大气局（National Oceanic and Atmospheric Administration, NOAA）和美国国家航空和宇宙航行局（National Aeronautics and Space Administration, NASA）开发的ETOPO1[①]和SRTM[②]数字高程模型，将这些基于栅格的地形数据与Macrostrat基于GIS环境的地理数据相匹配，用户可以通过移动应用程序或者网页界面进行访问。

2.7　Gplates模型

古地理环境对地球系统科学的众多问题具有重要意义，如重建气候敏感性沉积岩的时空分布（Cao et al., 2018）、研究大陆漂移对碳酸盐沉积的影响（Walker et al., 2002）、探索板块构造与生物多样性之间的联系（Zaffos et al., 2017）。因此Macrostrat为数据提供了基于GPlates平台的板块构造框架，可实现板块构造重建的交互式操作及各类数据在地质时间尺度上的可视化，并能够通过地球动力学计算将Macrostrat的各类数据与板块构造模型有效结合。Macrostrat数据与Gplates模型的结合是基于Python语言来实现的，其中，Macrostrat提供岩石地层单元的地质年龄和现代地理位置，Gplates提供相应古地理位置，目前只针对560 Ma以来的古地理重建。

2.8　系列产品

为了充分挖掘Macrostrat的现有数据，其团队开发了一系列快捷方便的网页端口或者移动端的软件工具，满足于各类用户需求。

（1）Macrostrat Beta

是Macrostrat专门用于数据访问的网页端口，目前已经更新至0.3版本。通过该端口，用户可以了解数据库当前的建设情况以及进行相关数据和文献下载；同时新开发的功能也将在该平台进行展示。

（2）Sift

是Macrostrat的搜索网络界面，是一款面向大众的可视化信息筛选器，目前可以根据时代、地层单元、岩性、地层柱、地层柱组、沉积环境、矿产类型对数据进行筛选。但Sift目前无法进行筛选条件的组合，如同时限定岩性和时代，另外也无法做到CSV源文件的导出。

对于有更精确的数据分析需求的用户，Macrostrat提供API（Application Program Interface应用程序界面）接口，通过API接口可以实现更自由的筛选数据，并获得CSV等格式的源数据。用户可以通过网页浏览器按照Macrostrat的预设规则，直接以API命令行形式访问数据库核心，筛选并导出数据。

（3）Map

是基于Macrostrat所收录的地质图开发的网络搜索界面，用于检索全球不同比例尺的地质图。

（4）Rockd

是Macrostrat团队开发的移动端APP，利用Macrostrat的API进行数据勘探和可视化，内部包括全球范围的地质图以及指向Macrostrat和

① ETOPO1：一种地形高程数据，包括陆地高程数据和海洋海底地形数据.

② SRTM（Shuttle Radar Topography Mission），即航天飞机雷达地形测绘任务，主要任务为获取地表雷达影像，绘制数字地形高程模型（百度百科）.

Geodeepdive 的链接。Rockd 用户可以轻松地记录实时地质现象，掌握实时考察的构造位置、地层概况，并使用实地的位置为附近的地质单元、化石提供空间信息建议。

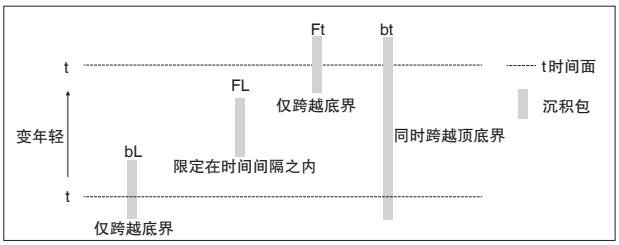
3 Macrostrat 数据库沉积物质的量化

解析地质记录的时空分布结构，需要获取以下量化数据：岩石数量、岩石类型、岩石地理、岩石沉积环境以及岩石记录的时间连续性。时间连续性指的是地质记录以一定的时空分辨率不间断地保存地质历史的程度。

3.1 量化的基本单元——Packages

Macrostrat 地层岩石量化的核心思想是：在地球表面的某特定位置的稳定沉积环境下，沉积物随时间流逝不断就位、沉积，直到稳定环境发生改变。Macrostrat 将形成于稳定沉积阶段的三维沉积体定义为一个沉积包（Packages）。沉积包之间发生沉积环境的变化，表现为两个方面，一是沉积停止甚至开始侵蚀，二是沉积物的性质发生变化，将这两种环境变化对应的阶段称为“间断”（gap）。

沉积包类似于由层序边界所限定的沉积体系域，不同之处在于层序地层界面是穿时的，而沉积包在时间轴上具有时间连续性。为了对地层柱进行量化，Macrostrat 类比古生物学描述物种时间跨度的方式，将一个被“间断”所约束的沉积包设想为一个生物分类单元（图 4），则给定任意的时间间隔，所有沉积包将归属于以下四种之一：沉积包仅跨越时间间隔底界（bL）、沉积包限定在时间间隔内（FL）、沉积包同时跨越时间间隔顶、底界



b 和 t 表示跨越区间的底部（bottom）和顶部（top）边界，F 和 L 表示在区间内的第一次（first）和最后一次（last）出现（据 Foote, 2000 改）

图 4 给定时间间隔内的沉积包类型

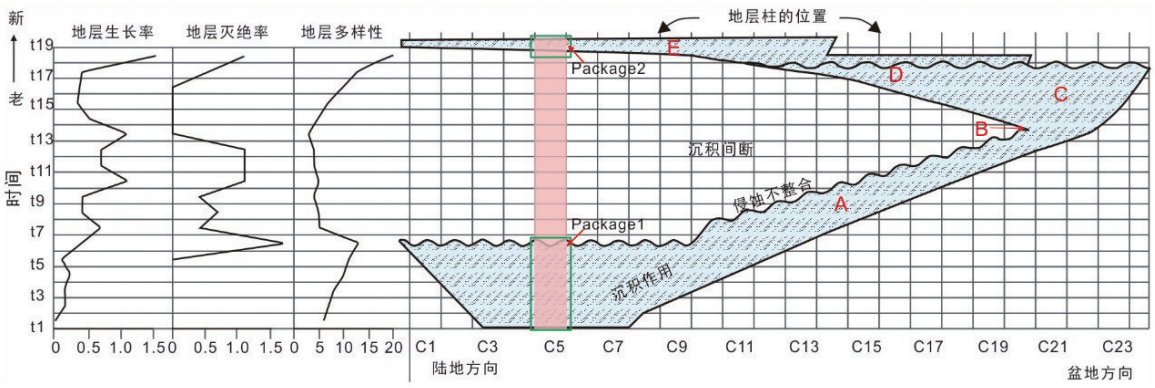
Fig. 4 Types of gap-bounded sediment packages present within a stratigraphic interval

（bt）、沉积包仅跨越了时间间隔顶界（Ft）。从而，地质记录可以借鉴古生物学的算法，计算时间轴上沉积包的“多样性”、“起源率”和“灭绝率”。

3.2 盆地尺度的沉积物质量化

地层柱代表了盆地的综合地质信息，模拟盆地尺度的量化是大陆尺度量化的基础。模拟的假设前提是在相邻时间间隔内的沉积包的持续分布概率遵循 Poisson 过程，即每个时间间隔内不同沉积包的发生是随机事件。通过统计时间轴上沉积包的类型和数量，即可对盆地的地层演化进行量化分析（图 5）。

（1）以单个地层柱为对象，确定纵向每一时间间隔内沉积包的类型。如图 5，红色阴影代表一地层柱（Column 5），由沉积间断划分为两个沉积包 Package1、Package2。在 t1–t6 的所有时间间隔内，Package1 均为 bt 类型沉积包；t6–t7 内，Package1 仅跨越了 t6，为 bL 类型沉积包；t7–t18 对应沉积间断；t18–t19 的顶部出现沉积，Package2 为 Ft 类型



t1–t19：时间间隔，C1–C23：地层柱，浅蓝色阴影部分：沉积作用，空白：沉积间断，A–E：各沉积阶段，详见正文（据 Peters, 2006）

图 5 理想化盆地尺度地层量化模型

Fig. 5 Schematic of stratigraphic quantification model at basin scale

沉积包；t19–t20 内，Package2 为 bt 类型沉积包；

（2）统计每一时间间隔内所有地层柱各类型沉积包的数量。如在 t1–t2 时间间隔内，仅 C3–C8 地层柱有沉积作用，C3、C8 表现为仅跨越顶界 t2 的 bL 类型沉积包，C4–C7 表现为同时跨越顶（t2）、底（t1）界的 bt 类型沉积包，即 X_{bl} =2，X_{bt} =4，X_{Ft} =0，X_{FL} =0（X 代表沉积包的数量）；

（3）根据经验公式计算各项量化指标：

$N = X_{bt} + X_{Ft} + X_{bl} + X_{FL}$，N 代表地层多样性，用于衡量盆地在某时间段内岩石沉积包多样性；

$p= -In [X_{bt} / (X_{bt} + X_{Ft})]$，p 代表地层起源率，用于衡量盆地在某时间段内岩石沉积包新生的速率；

$q= -In[X_{bt} /(X_{bt} + X_{bl})]$，q 代表地层灭绝率，用于衡量盆地在某时间段内岩石沉积包灭绝的速率。

（4）绘制演化曲线，解释定量化数据产生的曲线的地质学意义。如对图 5 的模拟可以得到以下结论：1）沉积地区收缩并快速向盆地移动时，形成不整合，对应地层多样性的大幅度脉冲（A）；2）当向盆地的沉积转变停止并且保存的沉积记录向空间扩张时，地层灭绝率下降为 0（B）；3）随着沉积区的扩张，地层多样性必然增加（C）；4）海侵使得沉积向陆转变，地层灭绝率和起源率都开始增加，即向陆的沉积作用提高了地层起源率，但是由于盆地内缺乏沉积物，地层灭绝率也相应提高（D）；5）最大洪泛面时对应最高的地层多样性（E）。

3.3　大陆尺度的沉积物质量化

整个北美大陆由多个沉积盆地组成，沉积盆地的地质信息由地层柱来表示，因此大陆尺度的量化将按照单个盆地依次处理，不同盆地的贡献将根据其面积进行加权。

地层综合柱状图反映了区域的地质信息，其具备了岩石种类、时代范围、厚度以及岩石地层单元、接触关系等属性，以国际地层委员会给出的地质年代为时间间隔，很容易提取每个时间间隔对应的沉积包类型及数量。Macrostrat 按照该方法人工编录统计了 COSUNA 和加拿大地质调查局（Geological Survey of Canada, GSC）显生宙所有地层柱的沉积包，并按照沉积环境或岩性对沉积包进行分类。

Macrostrat 通过以上大陆尺度的量化过程，获得初步量化数据：以“阶”（1~3 Ma）为时间间隔的不同类型、不同岩性、不同沉积环境的沉积包数量及其总量。以该数据为基础，Shanan E. Peters 团队对北美大陆显生宙沉积物的演化模式及相关科学问题进行了深入研究，将在第五部分进行详细论述。

3.4　面积和体积的提取

（1）Macrostrat 借助计算机技术为地层柱分配了地理多边形，每个多边形具有确定的面积（图 1）。根据地层柱给出的厚度，可以计算沉积物质的体积（Meyers and Peters, 2011）。

（2）借助于对地质图的解析来计算地层分布面积。由计算机地质制图得到的电子地质图，其岩石单元包括了一系列数字属性数据：面积、时代、岩石类型和名称信息等，因此可以通过直观的统计学手段得到各时间间隔内的不同种类岩石的面积分布。非电子版地质图，首先要对其进行数字扫描，利用图像分析软件将地质图转化为地理信息系统（GIS）格式，对图上每种岩石类型或每个岩石单元占据的像素计数，通过在每张地质图上的若干个 1°×1°的区域中，将累计像素缩放到真实区域，从而将其转化为大陆面积（Wilkinson et al., 2009）。

4　Macrostrat 文本挖掘技术

综合分析已发表的海量的文献数据，人工操作非常耗时，并且会生成一个与主要数据源断开连接的非扩展数据库。因此亟需建设一个可动态扩展的、可靠的网络基础设施，以促进发现、获取、利用和引用已发表文献中的数据和知识。

Macrostrat 除了提供开源的沉积学数据外，还提供了针对文献的机器阅读技术平台：Geodeepdive，即自动从已发表文献的文本、表格和图片中锁定并提取有用信息的技术。Geodeepdive 机器阅读主要涉及的计算机技术包括光学字符辨识、文档布局识别、自然语言处理和结构化查询语言。Geodeepdive 的目的是：（1）降低数据集成的时间和成本，将科学家的工作重心从缓慢且昂贵的数据整合工作转移到创造性的假设测试；（2）测试关键结论的重现性，加深对重大科学问题的理解；（3）促进机器阅读技术发展，尤其是在科研领域中得到部署和验证；（4）基于现有文献中的字段和样本，更集中、高效、智能地生成衍生数据。为了实现以上目的，Geodeepdive 与 8 大出版商（图 6）达成协议，获取巨大的文献数据库用于机器阅

读，且保证文献库中的原文保密，但数据公开。

Geodeepdive的工作模式分为3个步骤。第1步，科学家提出科学问题，确定需要挖掘的数据，然后使用Python、JavaScript、PostgreSQL等计算机语言写出算法，描述数据挖掘思路，即如何提取特征信息；第2步，使用超级计算机高速处理文献库的海量文献，按照预设算法进行挖掘，并生成因子图（用于表征各实体之间的关系）；第3步，输出机器挖掘的结构化数据和学习结果（图6）。通过机器阅读的工作模式我们可以发现，机器阅读或者文本挖掘过程是一个边工作边学习的过程，随着前提的改变或者新的数据的加入，产生的结果可能发生变化。同时，机器阅读系统能够利用非结构化的多源科学文献构建一个结构化的数据库。其中的数据都是具有概率的事实，整体上是一个与主要数据源紧密耦合的概率数据库，其数据质量可以与人工阅读和编译数据生成的数据库相媲美（Zhang et al., 2013; Peters et al., 2014a）。

例如,基于Geodeepdive的衍生工具,Paleodeepdive（PDD），主要服务于对化石数据的挖掘，用于加深对大规模生命演化史的理解，包括长期的分类多样性和基因组级灭绝和起源速率等问题的研究。通过PDD自动提取生物分类单元、地质岩层、地理位置和地质时间间隔等数据所建立的综合古生物数据库，在生物宏演化模式研究上获得了与人工汇编的PBDB相似的结果，因此有理由相信由机器阅读产生的结果是真实可信的。除此之外机器阅读更大优势在于，它生成的数据库类型与手动填充的数据库有本质的不同。在PDD生成的概率数据库中，每条数据都具有相应的准确性概率，且与其源文件中的上下文紧密耦合，甚至提供url链接。因此，只要对任何一个组件给出反馈，或者向系统添加额外的规则或数据，就可以系统地提高整个数据库的质量。更重要的是，PDD的数据采集过程是基于对整个文档的可视化和文本分析的，并且系统可以很容易地容纳更复杂的数据类型，例如生物插图中的形态学数据和相关的文本描述。因此，利用Paleodeepdive的系统能够识别和提取当前不属于数据库但与上下文相关的复杂数据（Peters et al.，2014b）。

Geodeepdive数字图书馆和机器阅读体系与Macrostrat平台相连，随时添加、编辑和发布新的地层、岩性、环境等数据，致力于用新的数据不断产出新的结果。

5 基于Macrostrat数据库的科学研究

Macrostrat收录了以北美地区为主的大量的地层和沉积学相关数据，但其核心价值不是体现在数据量的规模效应，而是基于数据相关性分析提供科学预测和假设（张旗和周永章，2017）。Macrostrat的首要目的就是帮助沉积学家解决全球尺度的大科学问题，如验证岩石保存和再旋回的地质假说、探索生物及生物化学演化的驱动力。

5.1 沉积物质循环

前人对地质历史时期沉积物质总量的循环规律主要有两种认识。

传统观点认为：由于侵蚀作用的累积，沉积岩总量必然随年龄增长而减少，并且具有指数衰减的趋势（Gregor, 1968）。该观点得到不同学者的进一步验证。Wilkinson等（2009）通过地质图面积提取发现沉积岩和火山岩的量具有随年龄增长

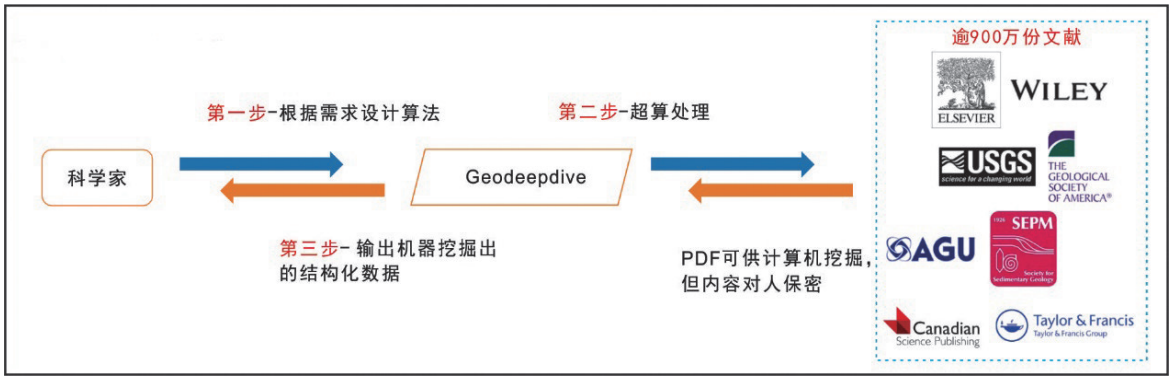


图6 Geodeepdive工作模式图

Fig. 6 Geodeepdive work pattern diagram

呈指数衰减的趋势，但是侵入岩和变质岩则无此趋势，其解释为不同的岩体形成于不同深度，接受到不同强度的侵蚀和埋藏作用。近来，Husson和Peters（2018）通过对埋藏速率和侵蚀速率进行模拟来观察保存岩石记录的演化趋势。其结果表明：无论埋藏和侵蚀速率是否是周期性或者在某一范围波动，只要侵蚀作用存在，岩石记录均随年龄增长而减小且趋于指数衰减。

第二种观点认为：大陆尺度下的沉积物总量是由总净沉积物累积速率决定的，并且具有周期性波动的规律（Ronov et al., 1980）。近年来，通过对地表不同年龄沉积物分布图像开展谱分析和回归分析发现：沉积物总量的演化周期接近56 Myr，与造山作用的周期相一致；显生宙沉积物的总量变化整体具有“M”形的演化趋势，与超大陆的旋回相关（Peters, 2008; Meyers and Peters, 2011）。

近年来，Shanan E. Peters团队采用大数据和地层定量化的方法对沉积物质循环问题开展了深入研究。该团队对显生宙不同岩性的沉积物进行量化处理之后发现：（1）在岩相组成方面，古生代沉积岩以碳酸盐为主，至新生代则几乎完全转变为陆源碎屑岩（图7），研究者将这种转变与劳伦大陆从低纬向高纬的移动联系起来；（2）沉积物总量在二叠纪—三叠纪之交表现出明显的脉冲（图7），将其解释为超大陆的旋回（Peters, 2006）。另外，Peters和Husson（2017）还基于不同的沉积环境对沉积物总量的演化曲线进行指数拟合（图8），结果表明：不同沉积环境的岩石具有不同的指数拟合程度，非海相和深海相沉积物的总量随着年龄增长呈指数降低，而浅海相沉积物具有多峰分布的特征。这是因为深海相沉积物只有在洋壳的某些部分形成，其破坏主要由俯冲控制，因此随着时间变老沉积物总量呈指数下降；非海相沉积环境下，侵蚀和岩石破坏作用是其主要控制因素，但沉积物所处的构造和环境极不均匀，导致非海相沉积物的指数匹配程度相对较差；对于浅海相环境，其沉积物分布面积广，数量大，成因多样，几乎可以在所有盆地的任何发育阶段进行大范围沉积，因此沉积物不随年龄变老呈指数降低（Husson and Peters, 2017; Peters and Husson, 2017）。

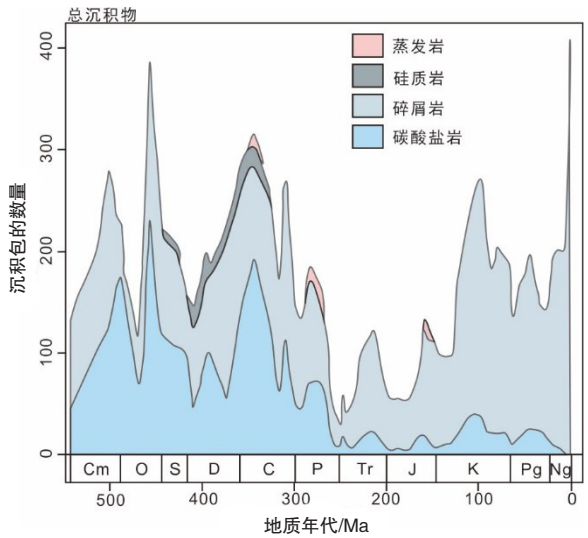


图7　显生宙北美地区沉积包随时间序列的变化图（据Peters, 2006）

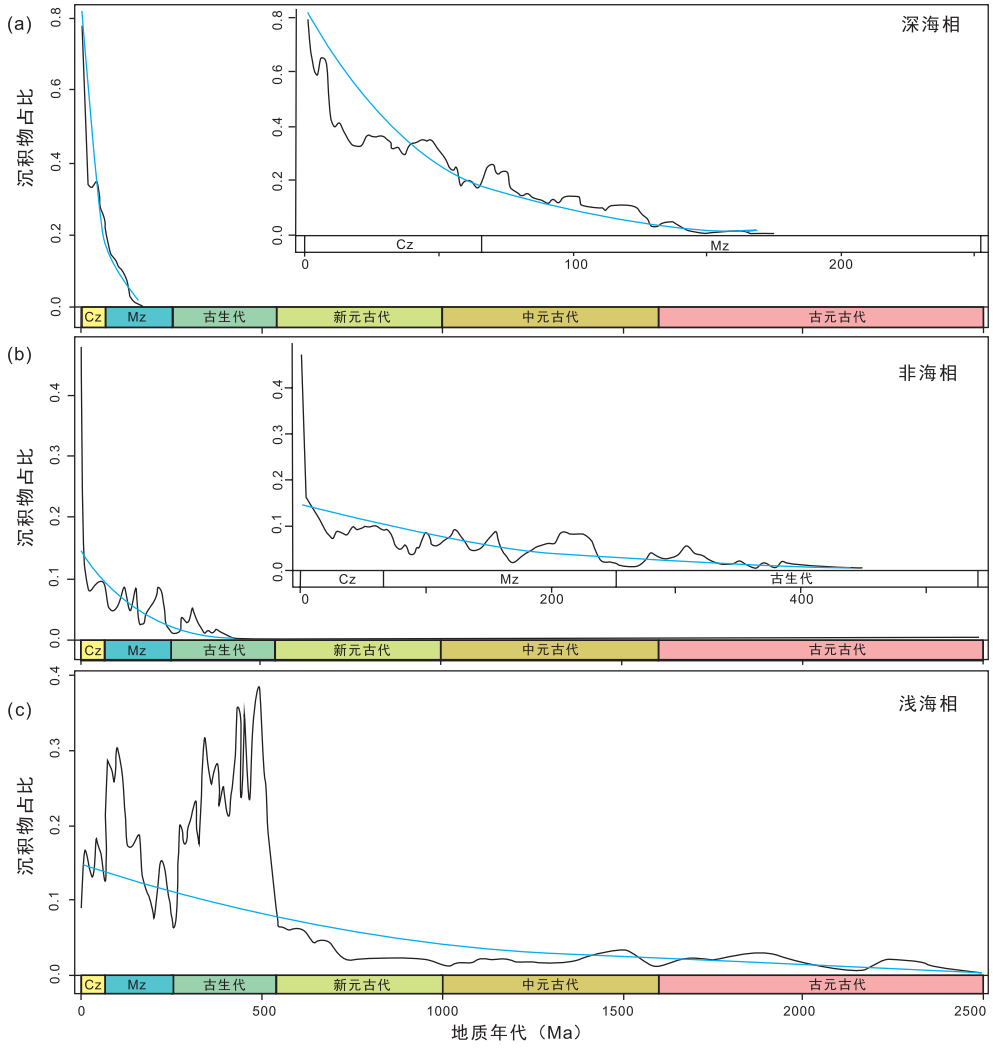
Fig. 7　Time series of the total number of sedimentary packages in North America at Phanerozoic (from Peters, 2006)

综上，沉积物质总量的演化主要受控于超大陆的旋回（Ronov et al., 1980; Peters, 2008; Meyers and Peters, 2011），侵蚀作用驱使沉积物总量随年龄增长而呈指数衰减（Wilkinson et al., 2009; Husson and Peters, 2018）；不同岩性的沉积物具有不同的沉积、侵蚀和埋藏条件；不同的沉积环境下，沉积物的沉积、保存以及演化模式也各不相同（Peters, 2006; Husson and Peters, 2017, 2018; Peters and Husson, 2017）。因此在研究沉积物质循环问题时，应对不同岩性、不同沉积环境的沉积物进行分别审视。

5.2　沉积物演化与生物宏演化的关系

宏演化（Macroevolution）指在物种层面或更高层次的进化，包括遗传学、形态学、分类学、生态学等上的变化（Mayr, 1982），与以基因演化、分子演化相关的微观演化相对应（Reznick and Ricklefs, 2009）。前人研究发现，现有的沉积岩记录与化石多样性之间存在相关性，这种相关性在海洋环境中尤为明显（Hannisdal and Peters, 2011）。因此，深刻理解沉积记录和化石记录之间的协变机制，对于理解生物多样性、物种起源、物种灭绝是至关重要的。

目前对于岩石—化石协变机制，学术界仍然存在争议。一方面，通过现有化石记录总结得到的多样性、起源和灭绝模式很可能是显生宙沉积岩记录中不完整的化石记录所导致的产物，即取



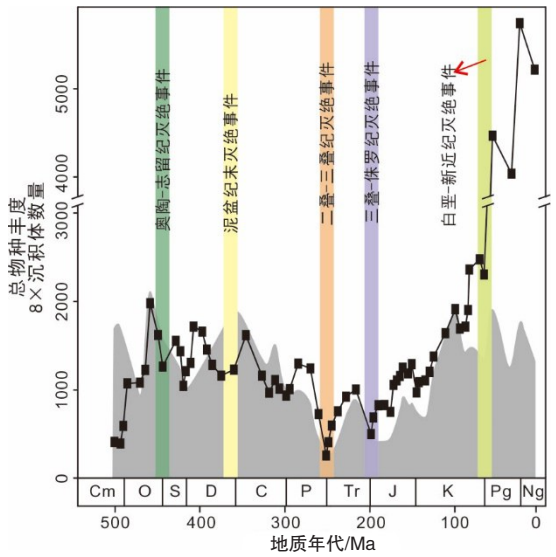
划分为三个构造和环境亚相（Cz：新生代；Mz：中生代）：(a) 深海相，(b) 非海相，(c) 浅海相；蓝色线为指数拟合曲线

图 8 Macrostrat 数据库沉积岩数量的时间序列演化图（据 Peters and Husson, 2017）

Fig. 8 Macrostrat database sedimentary rock quantity (Based on Peters and Husson, 2017)

样偏差导致（Peters and Foote, 2002; Mcgowan and Smith, 2008），一个明显的例子是地层不整合的出现严重影响古生物学家对生物多样性的评估（Peters and Foote, 2001, 2002），导致生物分类单元的人为聚类（Holland, 1995）；另一方面，尽管地质历史的生物多样性只能从不完整的岩石和化石记录中取样，但岩石记录的变化可能与生命的宏演化具有相同的控制因素，即一种共同的地质原因既决定了真实的灭绝速率，也决定了保存下来的沉积岩的数量（Heim and Peters, 2011; Peters and Heim, 2011）。

Macrostrat 数据库的沉积岩石记录和 PBDB 全球范围的化石记录 Peters and Mcclennen, 2016，为研究沉积物演化、生物宏演化及其协变机制提供了数据基础。PBDB 的化石记录可以与 Macrostrat 中的地层单元及其沉积环境相互匹配（Peters et al., 2018）。因此，以间断为边界、由沉积包组成的 Macrostrat 定量化数据可以用来检验取样偏差假说（Peters and Heim，2010）。近年来，Peters 和 Heim（2010, 2011）将北美沉积物和古生物演化数据进行对比发现：地层间断与物种起源或灭绝没有直接相关性；“地层起源率”和生物起源率之间亦没有强烈相关性；而“地层灭绝率”和生物灭绝率呈明显正相关，最突出的表现为沉积物质演化过程中大的沉积物间断与地质历史古生物大灭绝事件是相对应的（图 9），这种沉积物演化与生物起源和灭绝的不对称相关性表明岩石—化石协变机制不是由取样偏差决定的（Heim and Peters, 2011; Peters and Heim, 2010, 2011）。Peters 和 Heim（2011）进



Cm：寒武纪；O：奥陶纪；S：志留纪；D：泥盆纪；C：石炭纪；P：二叠纪；Tr：三叠纪；J：侏罗纪；K：白垩纪；Pg：古近纪；Ng：新近纪

图 9　总物种丰度（黑线）与沉积物总量（阴影）的时间序列演化图（据 Peters, 2005; Barnosky et al., 2011 改）

Fig. 9　Global genus richness (black line) and rock quantity (shaded area) plotted at age of interval base

(Revised after Peters, 2005; Barnosky et al., 2011)

一步研究发现，海洋生物的灭绝与海洋沉积区收缩期间发生的环境变化有因果关系，而海洋生物的起源与沉积区的扩张却没有呈现类似的关系，进一步验证了上述结论。

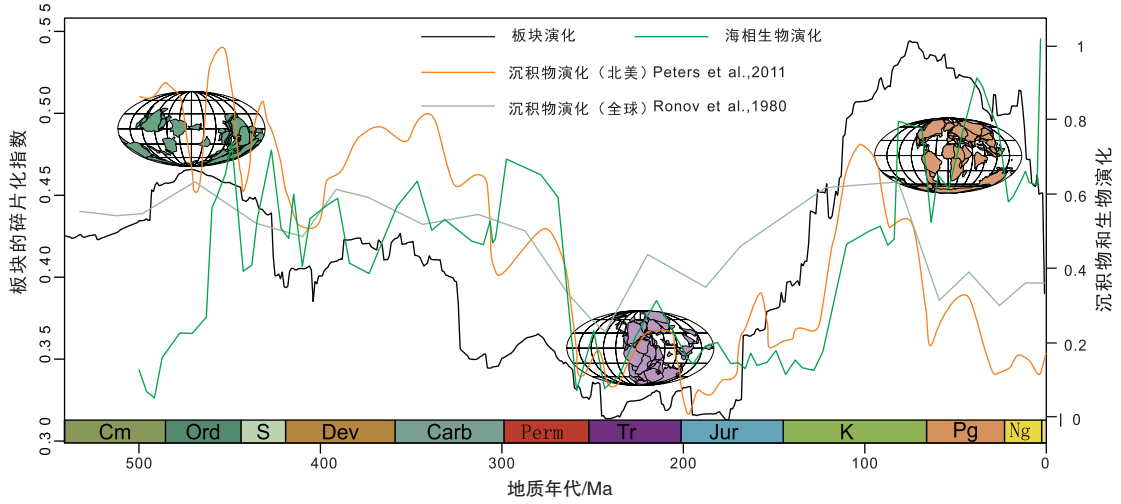
综上，前人对沉积和古生物大数据的对比研究表明：从生物灭绝的角度来看，沉积记录和化石记录的协变关系是地球系统之间直接或间接联系的综合记录；控制二者的共同机制可能涉及气候、构造、沉积和生物进化之间的众多直接和间接联系和反馈（Heim and Peters, 2011; Peters and Heim, 2010, 2011）。

5.3　沉积物质演化与地球系统演化

沉积记录的时空分布格局受多种地球过程（生物过程、构造过程、气候过程）控制，反过来沉积过程也在不同程度上改变和影响着地球过程（Hannisdal and Peters, 2010; Peters, 2008）。因此，在地质历史中得以保存的沉积岩是了解构造、气候和生命过程的重要档案。

（1）构造过程

沉积盆地的演化与大地构造演化密切相关，这是因为板块构造或者板块的相对位置控制着沉积盆地的类型（Dickinson, 1974; Ingersoll 1988; Busby and Ingersoll, 1995），区域的构造运动则通过控制对沉积物源区或沉积空间的形成和破坏来影响着区域的沉积记录（Peters, 2005; Meyers and Peters, 2011），因此，地质历史的沉积物与构造旋回往往同步演化，同时驱动生物演化（图 10）（Ronov et al., 1980; Zaffos et al., 2017; Peters and Heim, 2011）。



Cm：寒武纪；Ord：奥陶纪；S：志留纪；Dev：泥盆纪；Carb：石炭纪；Perm：二叠纪；Tr：三叠纪；Jur：侏罗纪；K：白垩纪；Pg：古近纪；Ng：新近纪

图 10　沉积物演化、海相生物演化（据 Hannisdal and Peters, 2011）与板块演化，板块的碎片化指数来源于以百万年为单位计算的 EarthByte 古地理重建模型（据 Zaffos et al., 2017 改）

Fig. 10　Sedimentary, marine biological and plate tectonic evolution, an index of continental block fragmentation derived from the EarthByte paleogeographic reconstruction models calculated in million-year increments

(Revised after Hannisdal and Peters, 2011; Zaffos et al., 2017)

（2）气候过程

气候过程主要通过驱动海平面变化影响区域和全球的盆地沉积过程（Miller et al., 2011; Meyers and Peters, 2011）；同时，冰期–间冰期的旋回也可作为沉积物类型的控制因素（Houten, 2000）。反过来，沉积过程可以通过掩埋和释放与气候变化相关的元素(主要是碳和硫)来调节全球气候，例如当前以碳酸盐或有机碳形式储存在沉积物中的碳远远超过了其它碳库，因此在某些时间尺度上，海洋–大气和地球表层之间的碳交换必然是推动气候变化的重要因素之一（Peters, 2005）。

（3）生物过程

生物过程通过多种方式（如生物扰动）直接影响沉积过程（Peters, 2005），如泥盆纪陆生植物的出现直接改变了冲积相泥质岩的比例（McMahon and Davies, 2018）。反过来，沉积过程通过影响环境来对生物过程的变化进行反馈，例如：生命和大气氧气历史上的主要特征就是通过定量描述保存沉积物总量随时间变化的幅度反映出来的（Peters et al., 2018）。近年来，Husson 和 Peters（2017, 2018）通过大数据对比研究发现：沉积岩的数量与地质历史氧气的变化以及生命的演化之间存在着强烈的过程联系，表明沉积岩的不稳定演化（有机碳相关氧还原、硅酸盐风化、洋壳沉积物的蚀变）驱动了氧气变化，进而驱动生命的演化（Husson and Peters, 2017, 2018）。

综上，前人的研究表明：复杂的构造过程、气候过程及生物过程共同决定了沉积物的时空分布特征；反过来，沉积物的形成过程也在积极地塑造地球系统（Hannisdal and Peters, 2010; Peters; 2005; Peters, 2008）。

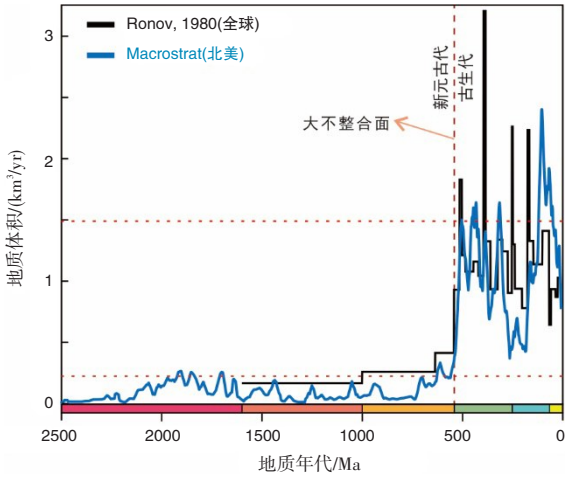
5.4 北美大陆大不整合面成因

地球上的沉积物直接盖在变质岩或岩浆岩等结晶基底之上，据全球各地观察，盖层和基底是截然接触的，二者中间存在一个侵蚀界面，代表时间间断，称为大不整合面（Great Unconformity）（Powell et al., 1875; Walcott, 1914; Yochelson, 2006; Karlstrom and Timmons, 2012）。导致大不整合面形成的成因争议很大，或与侵蚀基准面降低或者超大陆的聚合等因素有关（Sloss, 1963; Ronov et al., 1980）。近年来，Macrostrat沉积物定量化工作和地球化学数据库的建立为验证大不整合面成因提供了数据基础。

Macrostrat定量化沉积物体积的结果显示：在新元古代与古生代之交，沉积物体积增加了5倍之多，表明寒武纪之前大量的沉积物被侵蚀（Husson and Peters, 2018; 图11）。这一时期对应北美大不整合面的形成时期（Peters, 2006; Husson and Peters, 2017; Karlstrom and Timmons, 2012）。前人研究发现该时期陆地记录的地幔温度梯度和构造样式都没有明显变化（Keller and Schoene, 2012, 2018; Condie et al., 2016; Ganne and Feng, 2017），因此这种沉积响应与构造运动没有关系。最近，Keller等（2019）通过统计全球岩浆弧成因的锆石年龄、Hf和O同位素发现：εHf(t)在大不整合后降低、$\delta^{18}O$在大不整合后升高，这表明新元古代沉积物从陆壳消失而沉积在深海洋盆，进一步通过俯冲作用消减并改变了岩浆弧成分（Clift et al., 2009; Jagoutz et al., 2015）。Keller等（2019）进一步通过模拟方法对新元古代冰川侵蚀的沉积响应进行了定量化处理，发现3.4~4.5 km的冰川侵蚀量可以再现显生宙之前的侵蚀基准面。这一结果表明，新元古代“雪球地球”期间的冰川快速侵蚀是北美大不整合面形成的潜在驱动机制，同时也可能与不整合后寒武纪多细胞生命大爆发有直接或间接联系（Peters and Gaines, 2012）。

5.5 隐藏在大数据背后的非传统认知

大数据科学的特点之一是没有提前预设目标



Macrostrat数据库和Ronov的评估数据均显示元古宙与显生宙的沉积物体积相差5倍之多

图11 全球沉积物质体积演化（据Keller et al., 2019改）

Fig. 11 The evolution of global sedimentary rock volume (Revised after keller et al., 2019)

和前提，而是让数据“说话”，复杂多元的大数据所显示的内在关联，能够提高人类对经验世界的认知，这些认知往往出人意料（苏玉娟，2019）。

例 1：传统上认为叠层石的繁盛一般出现在生物大灭绝或生物多样性大幅降低之后，而其衰落则与生态稳定时期生物的持续演化和多样性增加相关（Schubert and Bottjer, 1992）。然而，近年来，Peters 等（2017）在利用机器阅读技术研究北美地区叠层石的时空分布的过程中，却得到了不一样的结论。为了从文献中智能提取叠层石数据，Peters 团队设计如下算法：在文献中搜索 Stromatolite（叠层石）及其衍生词汇，对包含这些词汇的句子及其相邻的句子进行自然语言处理，提取并记录叠层石词汇和岩石地层名称（视为潜在的产出叠层石的地层单元），在通过可靠性检验后，将文献、短语、地层属性等结构化信息反馈至用户。通过快速分析 8000 余篇相关文献，将其中出现叠层石的地层统计并投射至 Macrostrat 地层库中成图，结果表明：叠层石的出现与大灭绝并没有明显的相关性，而与白云岩总量的增长有很强的相关性（Peters et al., 2017）。

例 2：煤炭沉积是典型的气候敏感型沉积物，因此长期以来都被作为重建古纬度的有效工具（Diessel, 1992; Ziegler et al., 2003）。然而，近年来，Peters 等（2018）基于 Macrostrat 定量化的地层数据来验证上述问题时，得到不一样的结果。Peters 团队利用 Macrostrat 中全部包含煤炭沉积物丰度的相关数据，并使用 Matlab 内置函数将丰度量投射到时间序列之上；同时利用 Gplates 模拟煤炭沉积物的纬度分布，由此获得北美地区煤炭沉积物丰度随时间变化序列和煤炭沉积物的古纬度分布序列。结果表明：煤碳沉积物的古纬度分布在二叠纪初期明显向高纬度移动；二叠纪之后，煤的分布也并非恒定不变，其丰度和纬度都存在一定范围的波动。因此，Peters 等（2018）认为用煤炭沉积物重建古纬度时，其可靠程度有待进一步的验证。

例 3：通常认为沉积物通量对海平面的变化具有重要影响，反过来，海平面变化决定了进入沉积盆地的沉积物通量，进而控制了海相沉积物的总量（Ginsburg, 1982; Phillips and Slattery, 2006; Ferrier et al., 2015, 2019）。然而，近年来，Peters 团队通过大数据的整合分析发现：海平面对海相沉积物总量的控制不是通过沉积物通量的变化，而是与海平面变化导致的大陆洪泛面积的变化更为密切（图 12；Peters, 2008; Peters and Husson, 2017）。令人更难以置信的是，大数据分析表明大陆洪泛可以预测海洋生物的宏演化史（图 12；Peters, 2008），它们之间的相互关系表明：大陆洪泛面积可以作为一共同机制同时驱动海相沉积物演化和生物宏演化（Hannisdal and Peters, 2011）。

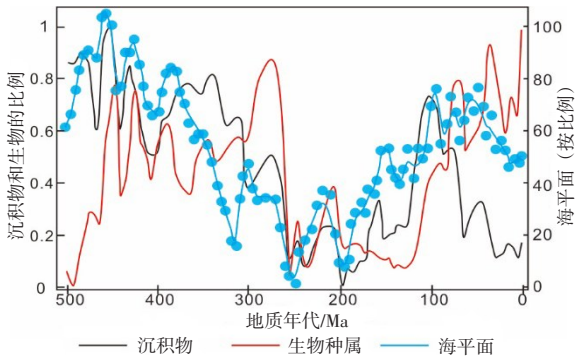


图 12　沉积物—生物种属—海平面显生宙变化图
（据 Hannisdal and Peters, 2011 改）

Fig. 12　Sediments, genera and sea level co-variation during the Phanerozoic (Revised after Hannisda and Peters, 2011)

6　结语

随着地球科学的发展、沉积学及相关数据快速增长，世界范围内涌现出一大批优秀的沉积数据库。早期的数据库以若干具体的科学问题为核心驱动，其建设、运营多依赖于少数科学家团队，尽管特点鲜明、专业性强，但是缺乏能动性和可持续性，并且在数据共享方面存在不足。Macrostrat 数据库是一个以岩石时空分布定量化为核心任务的跨学科数据平台，实现了在统一时空框架下对海量岩石、地层、生物资料的系统整合和定量分析，为深刻理解深时生命演化、地球物质循环、地质事件、古地理变迁、气候变化等提供了关键信息。然而，Macrostrat 数据库所产生的结论都源于其数据所覆盖的地理区域，这些结论放在全球尺度是否成立还需检验。另外，Macrostrat 数据库的核心数据基础是北美地层柱（Column）及其地层对比表。在高密度地层柱缺乏的世界其他地区如何开展此项工作是一个极大的挑战。

深时数字地球 (DDE) 计划建设开放、共享、

统一的大数据平台，将提供从全球尺度解决重大科学问题的契机。通过对沉积学领域内大数据整合和应用的深度调研，建议DDE大数据平台的沉积板块应当围绕沉积学的重大科学问题，有的放矢地进行数据的整合、分析、挖掘并进行预测；建立统一的时空框架和数据管理规则，高效整合复杂多元的沉积学数据；积极开发文本、图表信息挖掘技术，实现更加高效的机器阅读技术体系。

致谢：感谢评审人提出的细致而富有建设性的意见。

参考文献(References):

陈建平，李靖，谢帅，等. 2017. 中国地质大数据研究现状[J]. 地质学刊，41(3): 353-366.

陈友明. 1987. 全球沉积地质计划———一次国际学术会讨论报告[J]. 科学发展与研究（地球科学信息），2: 1-12.

姜浩端. 2013. 大数据的本质及其可能的影响[J]. 中国经济报告，6: 16-22.

刘宝珺. 1988. 全球沉积地质计划（GSGP）的制定和意义[J]. 四川地质学报，1: 44-50.

苏玉娟. 2019. 大数据知识的实现方法探析[J]. 山东科技大学学报（社会科学版），21(1): 20-26.

叶德燎. 1988. 全球沉积地质计划简介[J]. 地质科技情报，7(4): 92-94.

翟明国，杨树锋，陈宁华，等. 2018. 大数据时代：地质学的挑战与机遇[J]. 中国科学院院刊，(8): 825-831.

张旗，周永章. 2017. 大数据时代对科学研究方法的反思[J]. 矿物岩石地球化学通报，33(6): 881-885.

张维明，唐九阳. 2015. 大数据思维[J]. 指挥信息系统与技术，6(2): 1-4.

周永章，张良均，张奥多，等. 2016. 地球科学大数据挖掘与机器学习[M]. 中山大学出版社出版.

Barnosky A D, Matzke N, Tomiya S, et al. 2011. Has the Earth’s sixth mass extinction already arrived? [J]. Nature, 471: 51-57.

Brand L, Wang M M and Chadwick A. 2015. Global database of paleocurrent trends through the Phanerozoic and Precambrian [J]. Sci. Data, 2: 150025.

Boucot A J, Xu C, Scotese C R, et al. 2013. Phanerozoic paleoclimate: an atlas of lithologic indicators of climate [J]. SEPM Society for Sedimentary Geology, 11.

Busby C J and Ingersoll R V. 1995. Tectonics of Sedimentary Basins [M]. Blackwell Science.

Cao W C, Williams S, Flament N et al. 2018. Palaeolatitudinal distribution of lithologic indicators of climate in a palaeogeographic framework [J]. Geological Magazine, 156(2): 331-354.

Clift P D, Vannucchi P and Morgan J P. 2009. Crustal redistribution, crust-mantle recycling and Phanerozoic evolution of the continental crust [J]. Earth-Sci Rev., 97: 80-104.

Condie K C, Aster R C and van Hunen J. 2016. A great thermal divergence in the mantle beginning 2.5 Ga: Geochemical constraints from greenstone basalts and komatiites [J]. Geosci Front, 7: 543-553.

Dickinson W R. 1974. Plate tectonics and sedimentation. In tectonics and sedimentation [J]. SEPM (Society for Sedimentary Geology), Special Publication, 22: 1-27

Diessel C F K. 1992. Coal-Bearing Depositional Systems [M]. Berlin: Springer-Verlag, 72.

Dutkiewicz A, Müller R D, Callaghan S O’et al. 2015. Census of seafloor sediments in the world’s ocean [J]. Geology, G36883.1.

Ferrier K L, Mitrovica J X, Giosan L, et al. 2015. Sea-level responses to erosion and deposition of sediment in the Indus River basin and the Arabian Sea [J]. Earth and Planetary Science Letters, 416: 12-20.

Ferrier K L, Wal W V D, Ruetenik G A, et al. 2019. The importance of sediment in sea-level change [J]. Past Global Changes, 27(1): 24-25.

Foote M. 2000. Origination and extinction components of taxonomic diversity: general problems [J]. Paleobiology, 26(suppl.): 74-102.

Ganne J and Feng X. 2017. Primary magmas and mantle temperatures through time [J]. Geochem Geophys Geosyst, 18: 872-888.

Ginsburg R N. 1982. Actualistic depositional models for the Great American Bank (Cambro-Ordovician) [J]. Int. Congr. Sedimentol, 11: 114.

Ginsburg R N. 1986. Global sedimentary geology program [J]. SEPM Society for Sedimentary Geology, 1(5): 521-522.

Gregor C B. 1968. The rate of denudation in post-Algonkian time [J]. Royal Netherlands Acad. Sei. Proc., 71(1): 22-30.

Hannisdal B and Peters S E. 2010. On the Relationship between Macrostratigraphy and Geological Processes: Quantitative Information Capture and Sampling Robustness [J]. The Journal of Geology, 118(2): 111-130.

Hannisdal B and Peters S E. 2011. Phanerozoic Earth system evolution and marine biodiversity [J]. Science, 334(6059): 1121-1124.

Heim N A and Peters S E. 2011. Covariation in macrostratigraphic and macroevolutionary patterns in the marine record of North America [J]. Geological Society of America Bulletin, 123(3-4): 620-630.

Holland S M. 1995. The stratigraphic distribution of fossils [J]. Paleobiology, 21(1): 92-109.

Houten F B V. 2000. Ooidal ironstones and phosphorites-A comparison from a stratigrapher’s view [J]. Society for Sedimentary Geology, 66: 127-132.

Husson J M and Peters S E. 2017 Atmospheric oxygenation driven by unsteady growth of the continental sedimentary reservoir [J]. Earth & Planetary Science Letters, 460: 68-75.

Husson J M and Peters S E. 2018. Nature of the sedimentary rock record and its implications for Earth system evolution [J]. Emerging Top Life Sci., 2: 125-136.

Ingersoll R V.1988. Tectonics of sedimentary basins [J]. Geological Society of America Bulletin, 100: 1704-1719

Jagoutz O and Kelemen P B. 2015. Role of arc processes in the formation of continental crust [J]. Annu Rev Earth Planet Sci., 43: 363-404.

Karlstrom K E and Timmons J M. 2012. Many unconformities make one ‘Great Unconformity’ [J]. Grand Canyon Geology; Two Billion Years of Earth’s History, GSA Special Paper, 489: 73-79.

Keller C B, Husson J M Mitchell R N, et al. 2019. Neoproterozoic glacial origin of the great unconformity [J]. Proc. Natl. Acad. Sci. U.S.A. 116: 1136-1145.

Keller C B and Schoene B. 2012. Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago [J]. Nature, 485: 490-493.

Keller C B and Schoene B. 2018. Plate tectonics and continental basaltic geochemistry throughout Earth history [J]. Earth Planet Sci Lett, 481: 290-304.

Mayr E. 1982. Speciaition and macroevolution [J]. Evolution, 36(6): 1119-1132.

McGowan A J and Smith A B. 2008. Are global Phanerozoic marine diversity curves truly global? a study of the relationship between regional rock records and global Phanerozoic marine diversity [J]. Paleobiology, 34(1): 80-103.

Mcmahon W J and Davies N S. 2018. Evolution of alluvial mudrock forced by early land plants [J]. Science, 359(6379): 1022-1024.

Meyers S R and Peters S E. 2011. A 56-million-year rhythm in North American sedimentation during the Phanerozoic [J]. Earth & Planetary

Science Letters, 303(3–4): 0–180.

Miller K G, Mountain G S, Wright J D et al. 2011. A 180-million-year record of sea level and ice volume variations from continental margin and deep-sea isotopic records [J]. Oceanography, 24(2): 40–53.

Peters S E. 2005. Geologic constraints on the macroevolutionary history of marine animals [J]. Proceedings of the National Academy of Sciences of the United States of America, 102(35): 12326–12331.

Peters S E. 2006. Macrostratigraphy of North America [J]. Journal of Geology, 114(4): 391–412.

Peters S E. 2008. Macrostratigraphy and its promise for paleobiology [J]. Paleontol. Soc. Pap, 14: 205–231.

Peters S E, Husson J M and Czaplewski J. 2018. Macrostrat a platform for geological data integration and deep-time Earth crust research [J]. Geochemistry, Geophysics, Geosystems, 19: 1393–1409.

Peters, S E, Husson J M and Wilcots J. 2017. The rise and fall of stromatolites in shallow marine environments [J]. Geology, 45: 487–490.

Peters S E and Foote M. 2001. Biodiversity in the Phanerozoic: a reinterpretation [J]. Paleobiology, 27(4): 583–601.

Peters S E and Foote M. 2002. Determinants of extinction in the fossil record [J]. Nature, 416(6879): 420–424.

Peters S E and Gaines R R. 2012. Formation of the ‘Great Unconformity’ as a trigger for the Cambrian explosion [J]. Nature, 484: 363–366.

Peters S E and Heim N A. 2010. The geological completeness of paleontological sampling in North America [J]. Paleobiology, (1): 61–79.

Peters S E and Heim N A. 2011. Macrostratigraphy and macroevolution in marine environments: testing the common-cause hypothesis [J]. Geological Society of London, 358(1): 95–104.

Peters S E and Husson J M. 2017. Sediment cycling on continental and oceanic crust [J]. Geology, 45(4): 323–326.

Peters S E and Husson J M. 2018. We need a global comprehensive stratigraphic database here’sastart [J]. The Sedimentary Record, 16(1): 4–9.

Peters S E and Mcclennen M. 2016. The Paleobiology database application programming interface [J]. Paleobiology, 42(1): 1–7.

Peters S E, Zhang C, Livny M, et al. 2014a. A machine reading system for assembling synthetic Paleontological Databases [J]. Plos One, 9(12): e113523.

Peters S E, Zhang C, Livny M, et al. 2014b. A machine compiled macroevolutionary history of Phanerozoic life [J]. Computer Science, ArXiv Prepr ArXiv14062963.

Phillips J D and Slattery M C. 2006. Sediment storage, sea level, and sediment delivery to the ocean by coastal plain rivers [J]. Progress in Physical Geography, 30(4): 513–530.

Powell J W, Thompson A H, Coues E, et al. 1875. Exploration of the Colorado River of the West and its Tributaries [M]. Washington, DC: Government Printing Office.

Reznick D N and Ricklefs R E. 2009. “Darwin's bridge between microevolution and macroevolution” [J]. Nature, 457 (7231): 837–842.

Ronov A B, Khain V E, Balukhovsky A N, et al. 1980. Quantitative analysis of Phanerozoic sedimentation [J]. Sedimentary Geology, 25(4): 311–325.

Ronov A B, Migdisov A A and Barskaya N V. 1969. Tectonic cycles and regularities in the development of sedimentary rocks and paleogeographic environments of sedimentation of the russian platform (an approach to a quantitative study) [J]. Sedimentology, 13(3–4): 179–212.

Schubert J K and Bottjer D J. 1992. Early Triassic stromatolites as post–mass extinction disaster forms [J]. Geology, 20: 883–886.

Sloss L L. 1963. Sequences in the cratonic interior of North America [J]. Geological Society of America Bulletin, 74: 93–114.

Walcott C D. 1914. Cambrian Geology and Paleontology, Smithsonian Miscellaneous Collections [M]. The Lord Baltimore Press, Washington, DC.

Walker L J, Wilkinson B H and Ivany L C. 2002. Continental drift and phanerozoic carbonate accumulation in shallow-shelf and deepmarine settings [J]. The Journal of Geology, 110(1): 75–87.

Wilkinson B H, Mcelroy B J, Kesler S E, et al. 2009. Global geologic maps are tectonic speedometers-Rates of rock cycling from area-age frequencies [J]. Geological Society of America Bulletin, 121(5): 760–779.

Yochelson E L. 2006. The Lipalian interval: a forgotten, novel concept in the geologic column [J]. Earth Sci. Hist, 25: 251–269.

Zhang C, Govindaraju V, Borchardt J, et al. 2013. GeoDeepDive: Statistical inference using familiar data-processing languages [C] //ACM SIGMOD. International Conference on Management of Data. Association for Computing Machinery, New York, NY, USA: 993–996.

Zaffos A, Finnegan S and Peters S E. 2017. Plate tectonic regulation of global marine animal diversity [J]. Proceedings of the National Academy of Sciences, 114(22): 5653–5658.

Ziegler A M, Eshel G, McAllister R, et al. 2003. Tracing the tropics across land and sea: Permian to present [J]. Lethaia, 36: 227–254.