DEEP-TIME DIGITAL EARTH (DDE)

2020-2030

WHITE PAPER (Draft only)

International Union of Geological Science's
(IUGS) Recognized Big Science Program

March 2022

Authorship

DDE Science Committee

Roland Oberhänsli, IUGS / University of Potsdam, Chair
Danis Nurgaliev, Kazan State University
Dietmar Müller, University of Sydney
Hans J. Thybo, University of Oslo
Harsh K. Gupta, Geological Society of India
Guy M.Narbonne, Queen's University
Kathryn A. Whaler, University of Edinburgh
Kiyoshi Suyehiro, Tokyo University of Marine Science
Kombada Mhopjeni, Geological Survey of Namibia
Mónica Chamussa Juvane; Empresa Nacional de Hydrocabonetos
Muhammad Asif Khan, Karakorum International University
Nicholas Arndt, University Grenoble Alpes
Renata De Silva – Schmidt, Universidade Federal do Rio de Janeiro
Robert M. Hazen, Carnegie Institution
Shanan E. Peters, University of Wisconsin-Madison
William Cavazza, University of Bologna
Zengqian Hou, Chinese Academy of Geological Sciences

Editor in Chief

Chengshan Wang, IAS / University of Geoscience (Beijing)
Roland Oberhänsli, IUGS / University of Potsdam
Robert M. Hazen, Carnegie Institution

Writers & Contributors

Bin Luo, University of Geoscience (Beijing)
Chenghu Zhou (4), Institute of Geographic Sciences and Natural Resources Research, CAS
Feifei Zhang, Nanjing University
Hairong Lv (4.1), Tsinghua University
Haipeng LI, DDE RCE (Suzhou)
Juanle Wang (4.2), Institute of Geographic Sciences and Natural Resources Research, CAS
Junxuan Fan, Nanjing University
Michael H Stephenson (5, 6), Stephenson Geoscience Consultancy
Natarajan Ishwaran (6), Secretariat for IUGS DDE Program
Qiuming Cheng (4), University of Geoscience (Beijing)
Shaofeng Liu, AAPG / University of Geoscience (Wuhan)
Shihong Zhang, University of Geoscience (Wuhan)
Shuzhong Shen (3.2), Nanjing University
Susan Nash, AAPG
Tao Wang, Institute of Geology, Chinese Academy of Geological Sciences
Xiumian Hu (4.1), Nanjing University
Zhenhong Du (4.3), Zhejiang University

Liaisons

Chao Ma, Chengdu University of Technology
Haipeng Li, DDE RCE (Suzhou)

Reviewer

Boyan Brodaric, Geological Survey of Canada
Daniel Lebel, Geological Survey of Canada
Dietmar Müller, University of Sydney
Eric Boisvert, Geological Survey of Canada
Geneviève Marquis, Université de Montréal
Harsh Gupta, Cochin University of Science and Technology
Sally Pehrsson, Geological Survey of Canada
Sierd Cloetingh, Utrecht University

Editing & Design

Chao Li, Secretariat for IUGS DDE Program

Authorship Notes

- The authors' names are followed by markers indicating which chapter they were primarily involved in writing.
- The authors' names are listed in alphabetical order by first name.
- The coordinators included three people who were responsible for communication, collation and distribution during the writing and revision process.

Content

1. Letter to the Scientific Community

The Deep-time Digital Earth (DDE) is an innovative, international Big Science Program with the ambition to accumulate vast data resources related to our evolving, dynamic planet, coupled with advanced analytical and visualization methods, to facilitate a new era of data-driven discovery in Earth sciences. The DDE aims to provide new opportunities and directions for creative research and profound discovery in the study of Earth's storied past, its changeable present, and its uncertain future.

DDE was formally initiated as the first IUGS Big Science Program at an International Union of Geological Sciences (IUGS) event in early 2019 in Beijing, China, and has already benefited from international cooperation (IUGS, 2019, 2018; Wang et al., 2021). Leading experts from different fields, such as Earth science, planetary science, data science, and information science, have developed a set of groundbreaking pragmatic ideas and compelling aspirations. Based on those ideas and aspirations, this White Paper outlines the purpose of the DDE initiative, its science plan, and the strategy for its implementation.

The mission of DDE is to transform Earth science by harmonizing global geoscience data, sharing global geoscience knowledge, developing and disseminating advanced methods to analyze and visualize data, and fostering a deep-time data-driven research paradigm. Earth, upon which our lives depend, is a classic and challenging example of a macroscopic complex system. It has always been the long-term aim of Earth scientists to solve the mysteries of evolution and interactions among Earth materials, geodynamics, climate, and life. But in researching such a complex system, an enormous amount of complex data spanning Earth's immense volume over billions of years of planetary history must be assembled and manipulated. Much information has already been obtained, but an equal or greater quantity of data is not yet digitized or readily accessible. Furthermore, the acquisition of new data expands at an ever-increasing rate. To deal with these vast amounts of existing and new data ("Big Data"), we must think carefully about the most effective Earth science research strategy. The traditional paradigm of relying on small numbers of samples and limited data registered against present-day geographical coordinates cannot continue to provide the optimal approach to understanding the evolution and interconnectedness of the entire Earth system. Earth science must enter the age of Big Data.

To this ambitious end, DDE is dedicated to the harmonization of big and complex geoscience data around the globe to aid the development of a new deep-time, data-driven research paradigm. Information technology is now booming. DDE can be powered and carried along by that revolution. DDE will construct science infrastructures based on emerging technologies, including machine learning and artificial intelligence, with the aim of serving geoscientists around the world. The core purpose of these infrastructures will be to construct a machine-readable geoscience knowledge hierarchy for compiling and standardizing big and complex geological data from the Earth record. That bold vision will require the development of a world-leading data storage and analytics platform, so that geoscience questions of the highest societal and scientific importance can be tackled. For example, we are especially committed to supporting the realization of the United Nations 2030 Sustainable Development Goals. DDE will also strongly support interdisciplinary studies in fields including Earth science and information science, but extending to other relevant sciences, as well. A key principle will be to stimulate the establishment of International Research Centers of Excellence. Because of engagement with expert groups from different disciplines, there has been significant progress towards establishing a knowledge hierarchy covering all fields of solid Earth Science. A one-stop data platform and the first International Research Centre of Excellence will be established in Kunshan, China.

The long-term prospects for the ambitious DDE Program are much broader. Through harmonizing geoscience knowledge, data, and technologies, DDE will help achieve a clearer picture of ancient and contemporary Earth Systems, while providing vital insights into Earth's future.

2. Executive Summary

A major purpose of the International Union of Geological Sciences is to support a wide range of scientific studies on Earth systems through space and time, while advancing the application of results of scientific discoveries to protect Earth's natural environments, to use natural resources wisely, to alleviate impacts of natural hazards, and to improve the prosperity of nations and the quality of human life. Achieving a global understanding of the entire Earth system necessitates linking the solid Earth to the surface, hydrosphere, and atmosphere. But this goal is not possible without modern technology and data science. Therefore, to facilitate a transformation in the Earth sciences through data-driven discovery, the IUGS has proposed its first Big Science Program, the Deep-time Digital Earth (DDE), which is intended to operate from 2020 to 2030 (www.iugs.org/activities).

The overall **MISSION** of DDE is to harmonize and integrate deep-time Earth data, share global geoscience knowledge, and advance geoscience understanding and research. The long-term **VISION** of DDE is to transform the Earth sciences by fostering a deep-time data-driven research paradigm. Ultimately, DDE will help to identify better solutions to major global pure and applied research questions, such as the origin and evolution of life on Earth, the nature and driving factors in Earth's changeable near-surface environment, the distribution and recovery of natural resources, and the population capacity of Earth. DDE strives to identify, and distinguish among, natural and anthropogenic hazards; to characterize the self-organization of, and interactions among, complex Earth systems; and to predict the nature and consequences of global change in terms of the UN 2030 Sustainable Development Goals (SDGs).

The overarching **OBJECTIVES** of the DDE include enabling researchers to establish links among deep Earth processes; to characterize tectonic plates and their boundaries; to understand material transport and alteration at Earth's surface; to model whole-Earth energy and material flux budgets through time; and to develop a more robust understanding of Earth's interlinked physical, chemical, and biological mechanisms and feedbacks as a basis for predicting possible future evolutionary paths of the planet.

Fundamental scientific advances in understanding how Earth works will have important societal applications as well. For example, new technologies allow the rapid acquisition and manipulation of enormous amounts of data related to volcanic, seismic, weather, and other hazards, providing a means for significant progress in the risk analysis of contemporary Earth processes that occur infrequently. Analysis of the kinetics of these processes, which can have significant adverse impacts on society, is especially important. Such data have been available in well-curated and accessible databases such as EROS, ASDC, NOAA, UCAR, and EPOS[1], among others, for about 30 years. However, only young/recent and primarily near-surface processes can be characterized and modeled with modern monitoring techniques, using geophysical and geodesy surface and space-based sensor data. Well-structured, homogenized monitoring data are easily adapted to current analytical methods. However, unstructured and heterogeneous data, often referred to as "long-tail" data (e.g., Sinha et al., 2013), represent both challenges and opportunities in data analysis and visualization.

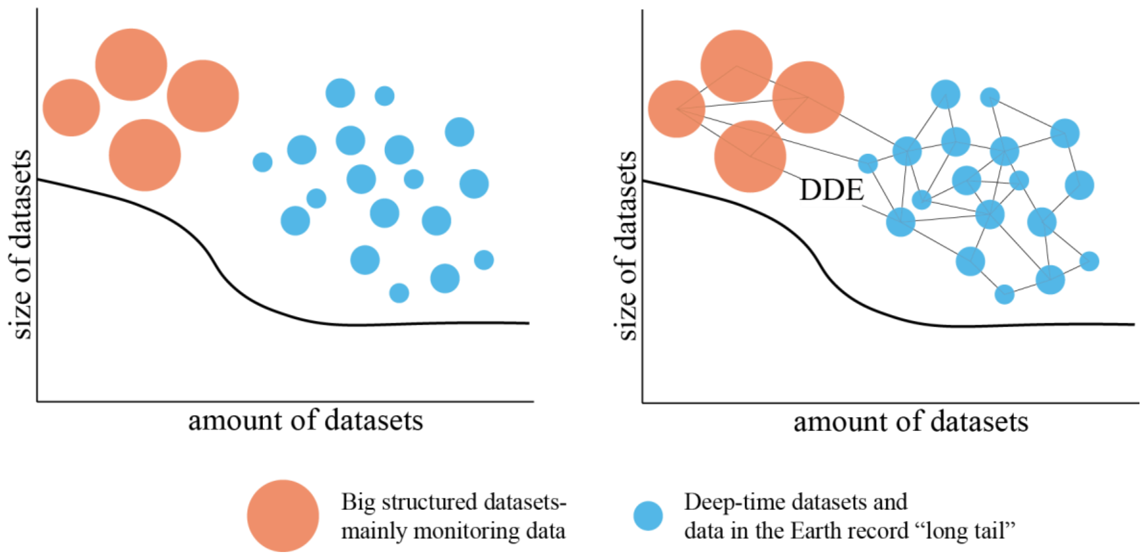

Fig. 2-1: New protocols, platforms and programs are needed to secure compatible and interoperable databases, so that the vast amounts of existing (and new) deep-time geoscience data can be linked. DDE will provide the "wiring" to connect deep-time data sources together.

Such long-tail or "dark" data related to the deep-time record of Earth processes are critical to understanding Earth's past and present state, yet much of that record is not readily available in digital form or has not yet been digitized. Such data have been generated in vast quantities by numerous individual scientists and scientific institutions over many years. These data, are often dispersed in varied local, often historical, databases and archives, and thus are difficult to access. There is an urgent need to make this material readily accessible in one system for geosciences to advance. Therefore, a major task of the DDE project will be to make long-tail, dark data accessible for improved investigation of the Earth system through time (Fig. 2-1).

The digital revolution makes it possible to harmonize data but several new tools will be needed to make databases compatible and interoperable, including new protocols, platforms, and programs.

[1] EROS: Earth Resources Observation and Science Center; ASDC: Atmospheric Science Data Center at NASA; NOAA: National Oceanographic and Atmospheric Administration; UCAR: University Corporation Atmospheric Research; EPOS: European Plate Observing System

Understanding Earth's 4.5-billion-year evolution requires innovation, using techniques such as "data mining", machine learning and artificial intelligence, Big Data analytics and visualization, and, ultimately, Internet cloud computing.

The proposed DDE Program and its vision were discussed at a workshop in Beijing attended by more than 70 scientists (Wang et al., 2019). They agreed that DDE should embrace the "FAIR" data concept – Findable, Accessible, Interoperable, and Reusable (Wilkinson et al., 2016)– enabling world-wide linking on desktop systems for geoscientists, students, and teachers.

DDE already involves many stakeholders including public institutions (i.e., research institutions, geological surveys, data repositories, and universities), learned societies, and professional associations. We anticipate that this list will expand significantly as the initiative progresses. This ambitious international program requires strong internal governance to ensure efficient decision making and transparent administration. An independent DDE Board will provide top-level governance of the Program. Key performance indicators (KPIs) will be set by a DDE Executive Committee. The Board will not make executive decisions nor will it be involved in the day-to-day management or running of DDE. A Governing Council (GC) will consist of one delegate nominated by each of the founding members of DDE and will oversee all activities of DDE. With the assistance of a DDE Secretariat, a DDE Executive Committee (EC) will bear the administrative responsibilities. A Science Committee will evaluate proposals submitted to DDE for approval by the GC. DDE Working Groups will undertake the core scientific activities of DDE giving special attention to infrastructure and knowledge systems of geoscience disciplines. DDE Task Groups will specifically explore databases through projects and/or workshops.

The main scientific goal of DDE is to achieve a better understanding of four main global issues related to planetary evolution:

I the evolution of Earth materials,
II the evolution of geography,
III the origins and evolution of life,
IV the evolution of climate.

In pursuit of these core scientific goals, the DDE Program has identified ten grand scientific challenges to provide better understanding of major global issues by analyzing Earth systems in the context of these four evolutionary aspects. Pursuing these questions will lead to a deeper understanding of the origin and evolution of life on Earth, the dynamics of the Earth-life systems, the prediction of global change, and the sustainability of our planetary home.

Better organized linking of data will transform Earth science so the initial practical challenge for DDE will be to insure the accessibility and interoperability of databases. Thus, a number of transformative innovations will be developed, including: a high-resolution integrative timeline for

Earth history; machine reading technologies for discovery and utilization of digital data and published documents; digitization and comparison of maps and graphic data; feature recognition in real-time field work; and evaluation and spatial-temporal association of Big Data.

We conclude that achieving these goals requires ambitious collaborations among research organizations, academic institutions, scientific associations, geological surveys, corporations, and governments. Several key issues that need to be addressed represent DDE's highest priorities:

1) A unified standard for all the data from different institutions and countries needs to be established, so that they can be compiled, compared, and analyzed within linked databases;
2) An interactive and collaborative software platform must be developed for processing, interpreting, and displaying the data and research results in 4-D; and
3) Artificial Intelligence technology must be adapted for the analysis of large quantity and multi-dimensional data to better understand complex Earth systems.

With the central goal of understanding Earth as a complex, evolving, living world though billions of years of history, and with the ambition of assembling vast, open-access data resources on the widest possible range of Earth systems to achieve this ambitious goal, DDE is poised to transform the study of our planetary home.

3. Illuminating the Future Through Earth's Evolution

"*All expectation for the future, indeed any idea that there will be a future, is conditioned on our understanding of the past*" [1]

As the world's population expands, prediction of the weather and climate disarsters, increasing demand for energy and resources, and ever-greater risks of geoharzards such as earthquakes and landslides call for deeper understanding of the interacting Earth spheres on different spatial and temporal scales. Billions of years of Earth history, including the evolution of material, geography, life, and climate, hosts the key for understanding current and future planetary change. The deep-time digital Earth data, from the ones hosted in well-curated databases to those dispersed in literature, thus form the basis for tackling the geoscience-themed "grand challenges". DDE recognizes the need to link and harmonize global deep-time Earth data, and embarks on a journey on data-driven abductive discovery at an unprecedented scale through international collaboration, setting the stage for a bright future of our planet.

3.1 Setting the Stage for the Future

The discipline of geoscience has reached the point where it must turn to recent great advances in gathering, storing, manipulating, analyzing, and visualizing vast quantities of relevant data. A suitable Big Data system and data-driven abductive discovery approach will enable us to solve both pure and applied questions in the interdisciplinary science of Earth – questions that were previously beyond our grasp.

More than half a century ago, the International Geophysical Year (1958) saw the birth of the World Data Center focused primarily on the geosciences, which has been transformed into a new World Data System (https://www.worlddatasystem.org/). Over the past three decades, progress in geoscience has led to improved access to large databases hosted at well-recognized data centers (Table 3-1).

Thanks to the foundation provided by these and many other large and expanding data resources, a transformation to a data-intensive approach to discovery in the Earth sciences has begun (National Academies of Sciences, Engineering, and Medicine, 2020). DDE, as an IUGS Big Science Program, has begun to identify pertinent science questions, while highlighting issues of data accessibility, linking, comparability, and the importance of cyber-infrastructure. We have embraced a community-based approach to assess needs and advances as well as to implement a

[1] quote in: Deep Time Earth-Life Observatory Network (DETELON) Science Plan Brochure (2011)

"FAIR" principle for cyber-infrastructure (Fig. 3-1). We aspire to go a step further by addressing both national and international standards for data acquisition, curation, and dissemination.

Table 3-1. Data organizations and Open-access data resources in the Earth Sciences

Database	Scope	URL and references
CODATA	CODATA exists to promote global collaboration to improve the availability and usability of data for all areas of research.	www.codata.org
World Data System (WDS)	WDS promotes long-term stewardship of, and universal and equitable access to, quality-assured scientific data and data services, products, and information across all disciplines.	www.worlddatasystem.org
EarthChem	Access to global geochemical and petrological data syntheses (PetDB, EarthChem Portal, LEPR, traceDs); EarthChem Library publishes and archives geochemical, petrological and mineralogical data as a trusted repository recommended by publishers	earthchem.org [1, 2]
Geobiodiversity Database (GBDB)	Integrated system for the management and analysis of section-based stratigraphic and paleontological information	geobiodiversity.com [3]
Macrostrat	Collaborative platform for the aggregation and distribution of geological data relevant to the spatial and temporal distribution of sedimentary, igneous and metamorphic rocks as well as data extracted from them	macrostrat.org [4]
Mindat	World's largest open database of minerals, rocks, meteorites and the localities they come from	mindat.org [5]
NASA Atmospheric Science Data Center (ASDC)	A leading provider of atmospheric science data products and services to the science community	asdc.larc.nasa.gov [6]
NOAA National Centers for Environmental Information (NCEI)	NCEI is responsible for hosting and providing access to one of the most significant archives on Earth, with comprehensive oceanic, atmospheric, and geophysical data	ngdc.noaa.gov [7]
OneGeology	Geologic map data and relevant geoscience data worldwide at scales≥1 : 1 million	portal.onegeology.org [8]
OneStratigraphy	Platform designed for sharing and using stratigraphic data, including integration, management, visualization and analytics of stratigraphic data	onestratigraphy.ddeworld.org [9]
Paleobiology Database (PBDB)	Global, collection-based occurrence and taxonomic data for organisms of all geological ages, as well as data services to allow easy access to data for independent development of analytical tools, visualization software and applications	paleobiodb.org [10]
USGS Earth Resources Observation and Science (EROS) Archive	World's largest collection of remotely sensed images of the Earth's land surface and the primary source of Landsat satellite images and data products	usgs.gov/centers/eros [11]
PANGAEA	An information system for processing, long-term storage, and publication of georeferenced data related to earth science fields	www.pangaea.de/[12]
Geochron	Collecting global sedimentary rock clastic mineral chronology data, dominated by clastic zircon age data	www.geochron.org/[13]
National Tibetan Plateau/Third Pole Environment Data Center	It possesses the most comprehensive scientific data on the Tibetan Plateau and surrounding regions of any data centers in China	data.tpdc.ac.cn/en/[14]
National Geoscience Data Centre (NGDC)	The geoscience data held at the NGDC includes a wide range of data types including but not limited to borehole, bedrock, hydrogeology, geochemistry, seismic, marine geoscience, oil and gas, airborne geophysical, and geohazards data	www.bgs.ac.uk/geological-data/national-geoscience-data-centre/[15]

[1-2] (Lehnert et al., 2019, 2000) [3](Fan et al., 2014) [4](Peters et al., 2018) [5](Hazen et al., 2019) [6](Ferebee et al., 2007) [7](Ansari et al., 2018) [8](Jackson, 2008) [9](Fan et al., 2020) [10](Peters and McClennen, 2016) [11](Loveland and Dwyer, 2012) [12](Diepenbroek et al., 2002) [13](Jiang et al., 2020) [14](Li et al., 2021) [15](Pinnick, 2017)

Fig. 3-1: "FAIR" data – Findable, Accessible, Interoperable, and Reusable – has become the new standard for building and sharing data resources (Wilkinson et al., 2016).

The current situation for geoscience data has some similarities to that of medical data in the late 20th century, prior to the time when large health-related data resources became widely interoperable and accessible. Many of those advances were made during the post-Second World War period, when computing began to be used to gather hitherto unavailable data on individual patient histories and lifestyles – multi-dimensional data that could be explored abductively to solve health and medical problems. Those multi-dimensional data, combined with new techniques of artificial intelligence, have brought enormous advances in medicine, for example, in documenting the roles of diet and exercise in health outcomes, in the diagnosis of chronic disease, and through machine recognition of cancerous cells in microscope slides. As extensive deep-time data resources become available to geoscientists, similar advances in understanding Earth's evolution can be anticipated.

Data useful for geoscience discovery can be considered in two categories (Fig. 3-2). One type of data consists of well-structured, homogenized "direct data" from field observations, laboratory investigations, and monitoring of solid Earth, hydrosphere, and atmosphere systems using sensor technologies. These data are usually already digitized, although some information is at risk if it is stored in outdated media and/or formats that may become unusable. Monitoring data supports the measurement, evaluation, and modeling of contemporary Earth processes, but those data are unable to elucidate past processes and events.

However, vast amounts of data generated by individual scientists and institutions are still not accessible and, in many cases, valuable data resources may not be known to the geoscience community. These sparse historical and legacy data have the potential to provide insights on the deep-time evolution of Earth, yet they remain out of reach. Significant amounts of so-called "dark" or "long-tail" data reside in poorly indexed and curated collections, often in obsolete formats that typically do not meet contemporary data recording standards (e.g., Heidorn, 2008; Sinha et al., 2013; Fig. 3-3).

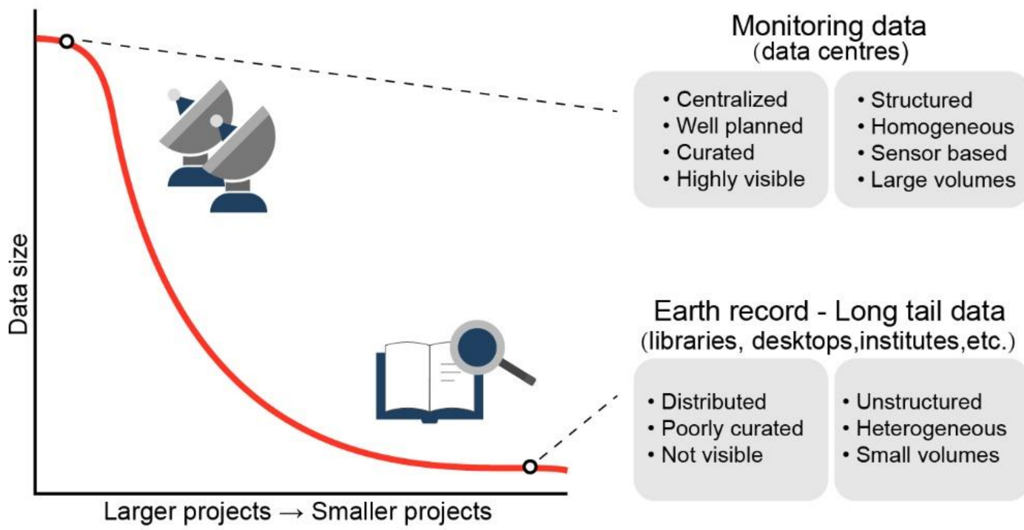

Fig. 3-2. One challenge of the DDE is gathering, organizing, and making available two contrasting types of geoscience data. Monitoring data are typically well-structured and interoperable. By contrast, "long-tail data" are often not accessible, residing in poorly indexed and curated collections, often in obsolete formats that typically do not meet contemporary data recording standards.

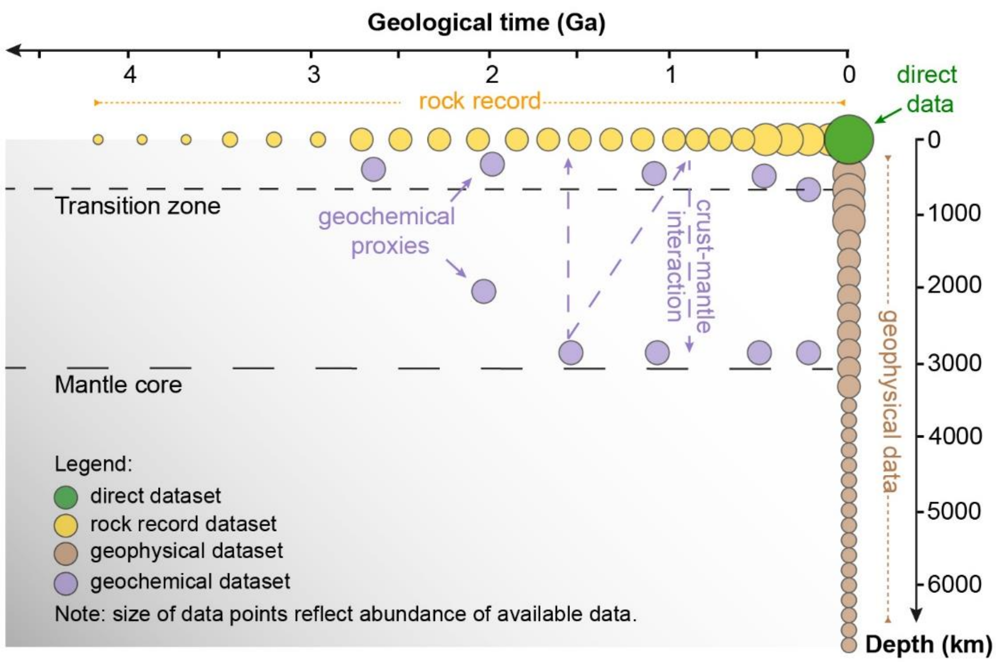

Fig. 3-3. Schematic time–depth diagram outlining data sets available to support research on Earth evolution. Size and spacing of data points reflects availability of data. The volume of geological data decreases with depth within Earth, as well as with greater age, in conformity with the "long-tail" concept of geodata.

DDE's vision is to transform geoscience by connecting, harmonizing, and making readily available the geological record of long-tail, deep-time data "islands" to support broad-based scientific studies relevant to the entire Earth system. The results of these and other studies will help us achieve a more comprehensive understanding of Earth's evolving natural environment and help in the wise use of natural resources for the prosperity of nations and the quality of human life.

The main approach adopted by DDE to investigate is what we call data-driven abductive discovery (Fig. 3-4; Box 2). This is made possible by the growing availability of vast, but largely untapped, Earth and life science data resources.

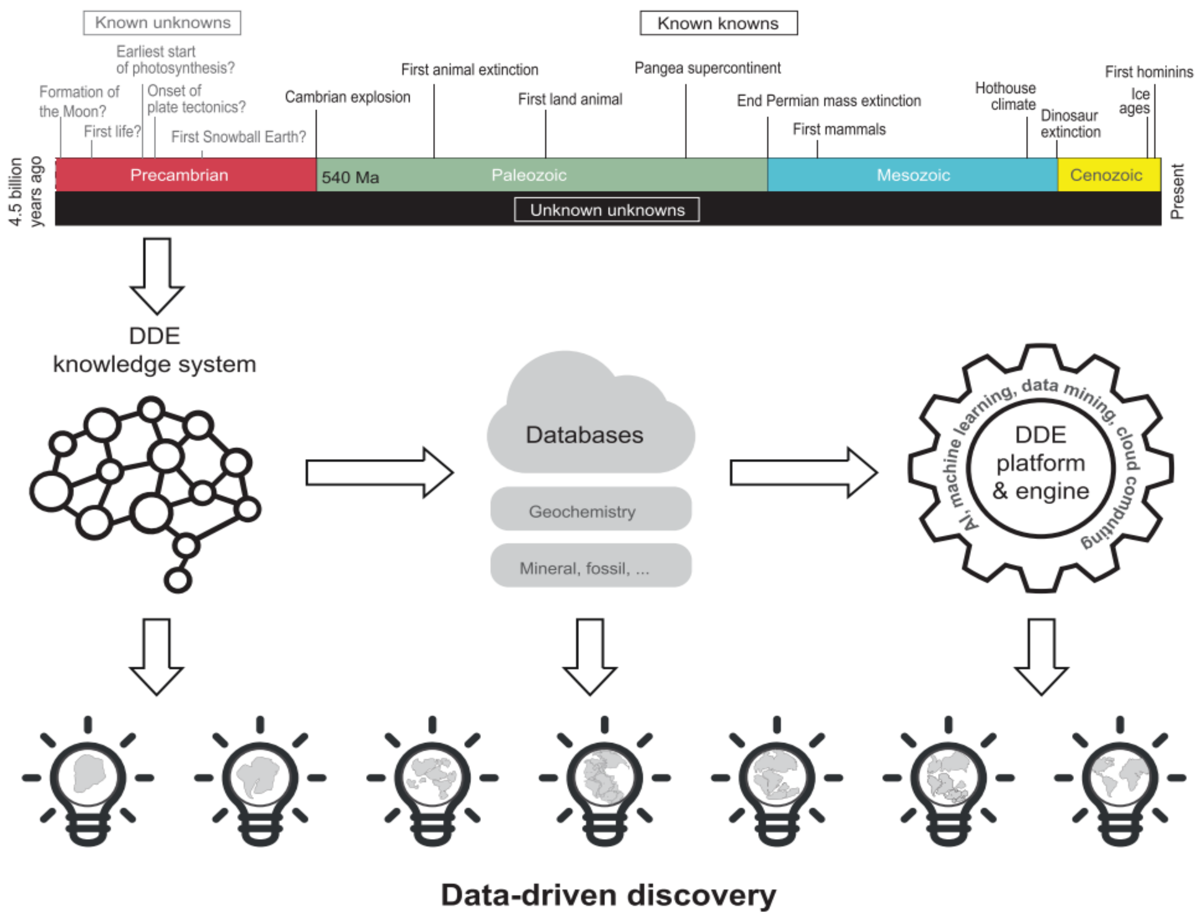

Fig. 3-4. Methods of data-driven discovery employ large, multi-dimensional Data resources and a Platform of advanced analytical and visualization methods to reveal previously hidden patterns in the data – a process of Data-Driven Discovery known as "abduction" that leads to New Hypotheses.

The DDE therefore advocates a strategic, data-driven program for accelerated scientific discovery. The transformative DDE initiative will enable scientists around the world to harness the power of data-driven discovery, thus advancing our ability to tackle major global issues, such as:

- integrating a uniform, high-resolution timeline of Earth history;

- refining our understanding of life's origin, evolution, and shifting biodiversity;
- elucidating the nature and evolution of sedimentary cycles;
- reconstructing and modelling the history of Earth's climate and atmosphere;
- documenting global sea-level change through deep time;
- quantifying the history of plate tectonics and deformation in four dimensions;
- probing the 4D architecture and evolution of deep-Earth materials and dynamics;
- assessing the evolving diversity and distribution of minerals through deep time;
- establishing a globally shared energy resource data infrastructure for sustainable development;
- creating a system of geophysical fields for prediction of seismic hazards.

The ambitious DDE Project thus strives to set the stage for a future of accelerated data-driven discovery in all facets of the Earth sciences.

It has been recognized that the discoveries and interactive usages of data, knowledge, and technology are of vital importance yet difficult for studying Earth, a hugely complex system. Therefore, there is an urgent need for an open, easily accessible platform for scientists worldwide to collaborate and increase the efficiency of discovering and using data, knowledge, and technology. DDE aims to fill this gap and is committed to building a multi-language cyberinfrastructure and an online community to promote the application of artificial intelligence, machine learning, cloud computing, and data management in the geoscience research field. Every person or team can solve and discover scientific questions in a secure space with free and powerful tools in this cyberinfrastructure. At the same time, the cyberinfrastructure and the online community encourage scientists to work as interdisciplinary teams to explore the potential of emerging information technologies in solving geoscientific problems. And DDE regards this cyberinfrastructure as one of the key responses to UNESCO's open science recommendations and will continuously improve its openness, global accessibility, and inclusion.

3.2 Data-Driven Discovery: Deciphering Earth's Evolution

Humans have beening pondering upon the three big questions since the dawn of civilization: the evolution of the universe, the evolution of Earth and the evolution of life through deep time (Box 1). Geoscientists have embraced the mission of elucidating the evolution of Earth and life, which are preserved in the information-rich but incomplete geological record that spans more than 4.5 billion years of Earth history. The most fundamental unanswered questions about Earth focus on its ongoing evolution as a dynamic planet: the evolution of Earth materials, of geography, of climate, and of life. Yet a comprehensive understanding of Earth's evoultion is hindered by the vastness of our planet's spatial and temporal dimensions – a surface world of more than 500 million square kilometers that has been evolving for more than 4.5 billion years. How can we possibly decipher that epic story of Earth? The new data-driven strategy adopted by DDE may hold the key.

3.2.1 The Evolution of Materials

The materials that form our planet evolved through the action of physical, chemical, and ultimately biological processes over more than 4.5 billion years. A central scientific objective of the Deep-time Digital Earth program is to document and understand our planet's material evolution, from crust to core throughout that long history. The evolution of minerals and sediments exemplify DDE's efforts to document these profound changes through deep time.

3.2.1.1 Mineral evolution

Minerals are natural solids that preserve Earth history for billions of years. Large and growing data resources on the diversity, distribution, and properties of minerals provide the foundation for a new era of data-driven discovery in mineralogy and, by extension, elucidating chemical and physical processes that have transformed our planet. Comprehensive international mineral databases include information on more than 5700 approved mineral species – their properties, ages of formation, localities, and more (https://mindat.org**;** https://rruff.org**)**. We analyze and visualize these data using diverse techniques that lead to a greater understanding of the co-evolving geosphere and biosphere as manifest in the diversity and distribution of Earth's solid materials (Hazen et al., 2019).

New data-driven directions include:

1) "mineral evolution," which explores the diversity and distribution of minerals in Earth, as well as other planets and moons, through deep time;
2) "mineral ecology," which analyzes the diversity and distribution of minerals across the globe and employs statistical methods to predict Earth's "missing" mineral species—those that exist but have not yet been discovered and described;
3) mineral network analysis, in which coexisting suites of minerals are visualized using the same methodologies as social network modeling (Fig. 3-5);
4) association analysis, by which previously unknown valuable mineral deposits can be located by examining positive and negative correlations among mineral associations. These methods, collectively, consider the distribution and diversity of minerals through space and time. They also foster a deeper understanding of mineral co-occurrences and can, for the first time, facilitate predictions of undiscovered mineral species.

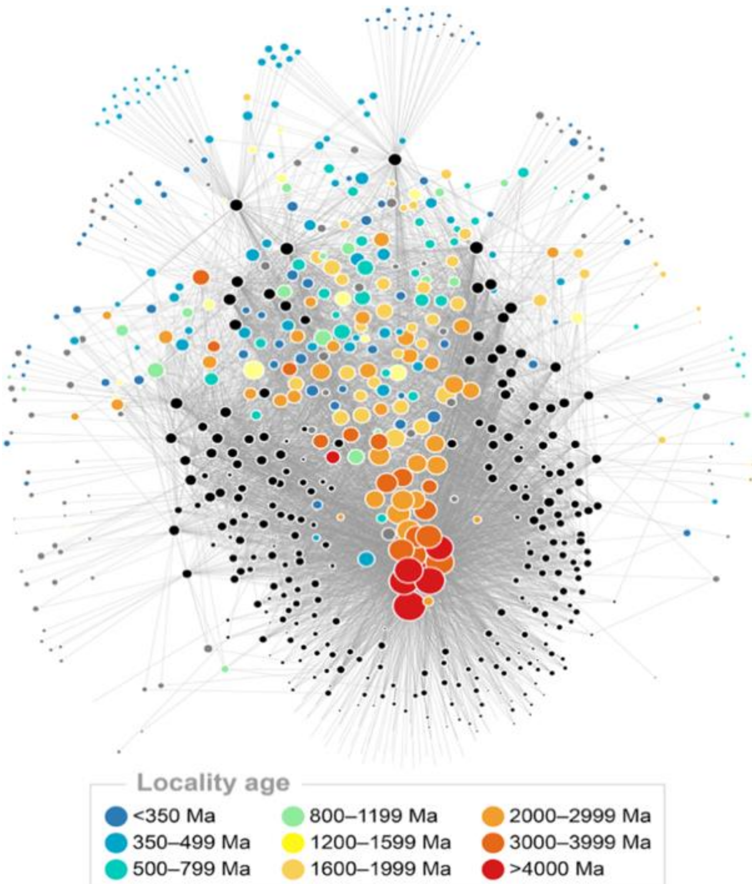

Fig. 3-5. Mineral Network Analysis: A bipartite network of 403 carbon-bearing mineral species (see also http://dtdi.carnegiescience.edu/node/4557 for an interactive version). Colored circles represent carbon mineral species. Circle sizes represent relative frequency of occurrence and colors (see inset scale) show the ages of the earliest known occurrences of those minerals. Black circles represent regional localities, with sizes corresponding to the relative numbers of different carbon-bearing minerals found at those localities. The network reveals important information regarding the diversity and distribution of carbon minerals through space and time. In particular, the "U-shaped" distribution of the black locality nodes, with only a few very common carbon minerals "inside" and many rarer carbon minerals "outside," is a visual representation of mineral distributions on Earth. Note that most of the commoner minerals are more ancient whereas most of the rarer minerals are more recent (after Hazen et al., 2019).

Some key questions are:

1) Can we predict as yet undiscovered minerals and where they might be found?
2) Can we use mineral data to better understand the evolution of Earth's geography, life, and climate?
3) Can the data resources that inform mineral evolution and mineral ecology provide a framework for the discovery of economically important mineral resources?

3.2.1.2 The evolution of sediments

Sedimentary rocks and unconsolidated sediments cycle dynamically and have evolved in deep time. They provide important information on the evolution of Earth's lithosphere, biosphere, and hydrosphere. Fundamental parameters, such as the total amounts, composition, flux, and spatial-temporal distribution of sediments, are used to elucidate tectonic, climate, and geodynamic processes in geological time (Fig. 3-6).

The DDE will systematically link existing sedimentary databases and facilitate analysis of information on deep-time sedimentary rocks (e.g., ages, regions, lithology, geochemical composition, rock types, paleoenvironments, tectonic settings), and quantitatively depict the spatial-temporal distribution of deep-time sediments. Big Data analytics will determine the evolutionary laws of the distribution pattern and reveal implications for Earth's tectonic evolution, life, sea level, and climate.

Some key questions are:

1) What is the spatial-temporal distribution of global sedimentary rocks?
2) What is the spatial-temporal distribution of special sedimentary rocks, such as bio-reefs, microbial rocks, dolomite, salt, coal, and oil shale?
3) Can the sedimentary architecture of continental clastic sedimentary rocks be reconstructed from source to sink through deep-time?
4) What are relationships among sediments and biological, tectonic, and climatic evolution?

3.2.2 The Evolution of Geography

Earth is a dynamic planet, with transfers of matter and energy that have affected the crust, mantle, and core in profound ways through 4.5 billion years of change. A fundamental goal of the DDE is to exploit large and growing data resources to deduce the evolution of Earth's geography, including variable sea level and the distribution of continents.

3.2.2.1 The evolution of sea-level

Currently, global sea-level is rising at a rate of close to 4 mm/yr (IPCC, 2021), with increasingly profound impacts on human habitats. Most models suggest that ocean thermal expansion and glacier melting are the dominant contributors to that rise, but there is no clear consensus on how sea level will change if global temperature continues to rise. The chemical and physical properties of sedimentary rocks record sea-level changes under different climatic states. However, no one method can completely reveal the global signal by itself.

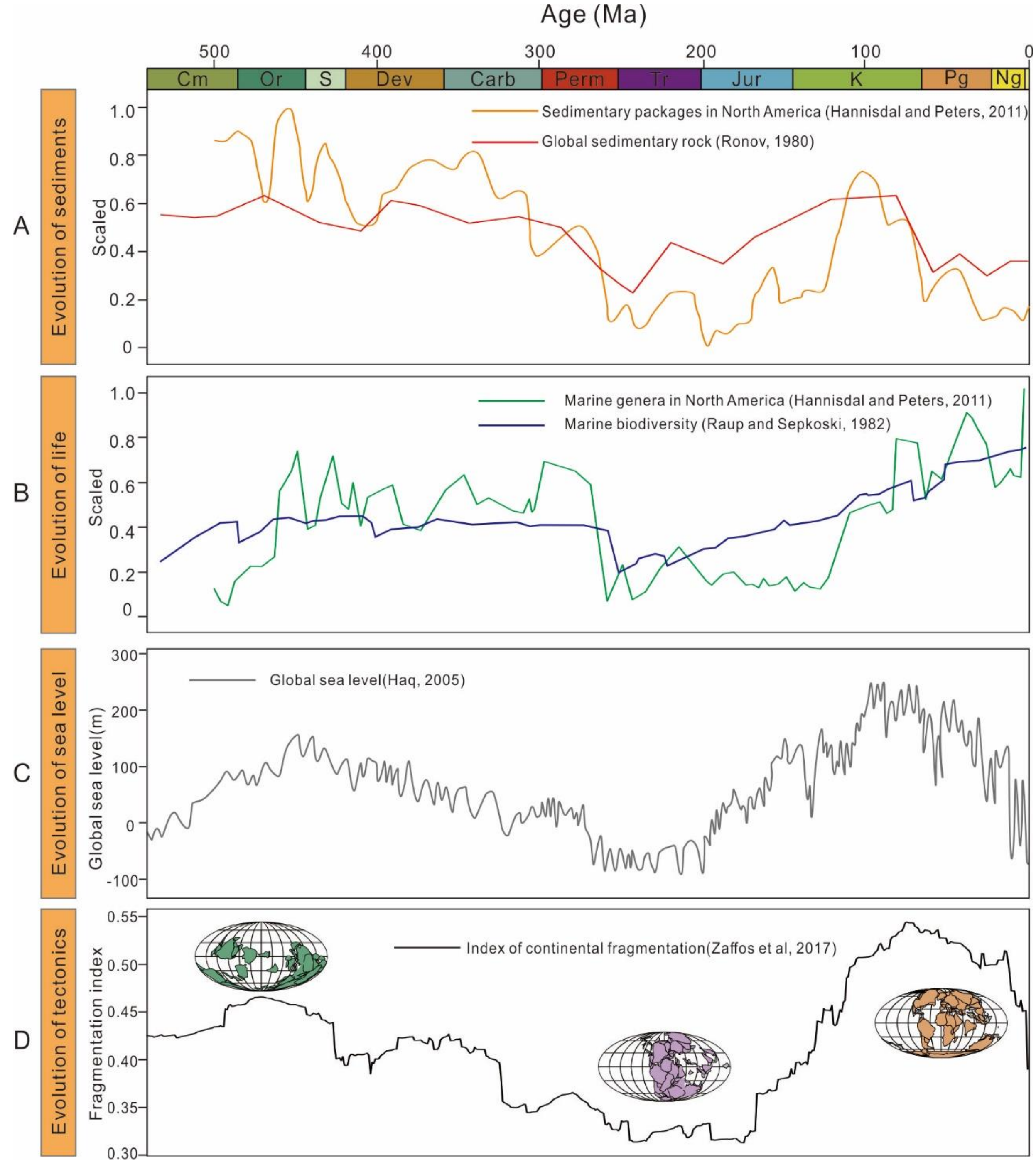

Fig.3-6. The deep-time evolution of sedimentary materials has been closely correlated with changes in animal diversity, sea level, and tectonics over the past 550 million years. A: The extent of sediments deposited globally and in North America (Hannisdal and Peters, 2011; Ronov et al., 1980); B: The total global marine biodiversity and North American marine genera (Hannisdal and Peters, 2011; Raup and Sepkoski, 1982); C: Sea level (Haq and Al-Qahtani, 2005); D: An index of continental block fragmentation (Zaffos et al., 2017). The index of continental block fragmentation derived from the EarthByte paleogeographic reconstruction models calculated in million-year increments. An index value of unity indicates no plates are touching and an index value of zero indicates all of continental blocks were contiguous and touching each other.

A more reliable way to reconstruct sea-level changes through deep-time is to combine studies based on different methods, synthesizing them into a uniform timeline for Earth history. The high-resolution "chronostratigraphy framework" to be constructed by the DDE, coupled with datasets from sedimentology, paleobiology, and plate tectonics, will enable a comprehensive synthesis of sea-level history on different time scales and under different climate states (Fig. 3-7).

Some key questions are:

1) How different have the rates and magnitudes of sea-level changed in deep-time?
2) What are the controlling factors of sea-level changes at different time scales and climate state?

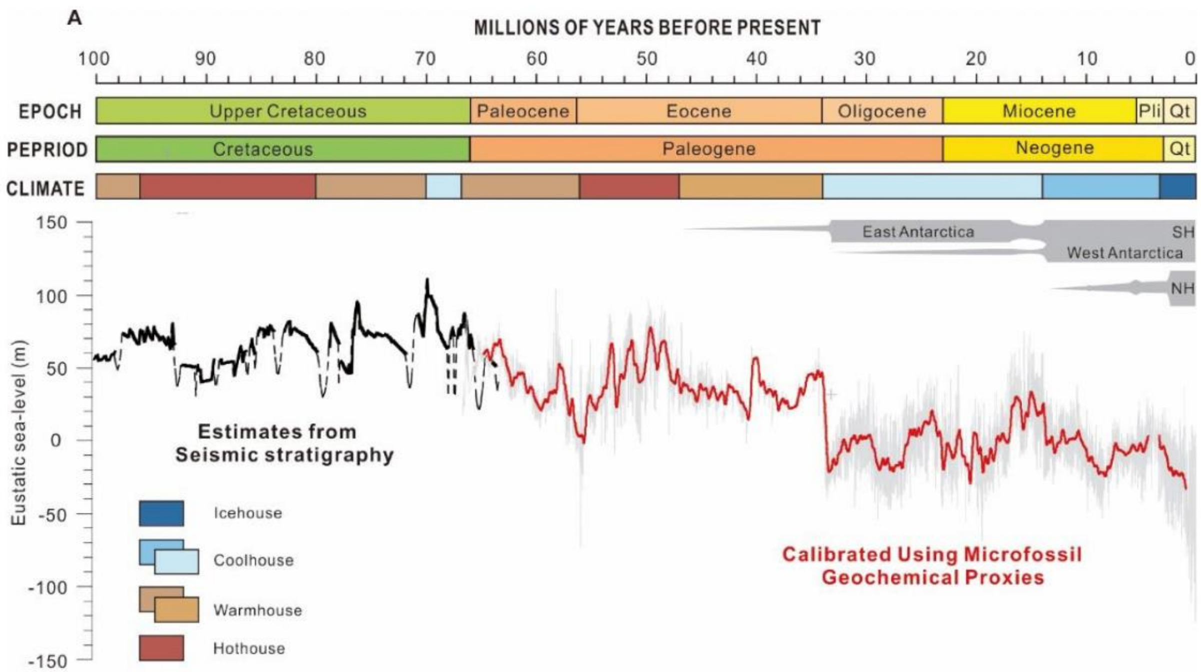

Fig. 3-7. The average sea-level of Earth's oceans, compared to modern values, have varied significantly over the past 100 million years (Miller et al., 2020, 2005). These changes that are tied to a variety of climatic and tectonic factors.

3.3.2.2 The evolution of plate tectonics

Plate tectonics, a unifying concept in Earth sciences, posits that Earth's surface is formed from a dozen dynamic "plates," which may be thousands of kilometers wide but average only a few tens of kilometers thick. Over hundreds of millions of years, these tectonic plates shift positions, interact with adjacent plates, and change in shape and extent. New crust forms at volcanoes along divergent plate boundaries, while old crust sinks into Earth's mantle at convergent plate boundaries. As plates diverge, oceans form; when plates collide, mountain ranges may rise up. Understanding the nature and history of plate tectonics is thus fundamental to documenting the evolution of Earth's geography, because tectonic forces directly control much of Earth's topography, while having significant influences on the evolution of climate and life.

Two simplifying assumptions of the plate tectonic paradigm have been that tectonic plates are rigid and that they are separated by narrow boundaries. However, many plate boundaries are characterized by deformation over significant areas; i.e., plates are not entirely rigid, nor are plate boundaries always narrow. As diverse data to quantify plate tectonics accumulate, the nature of tectonic deformations, and the feedbacks between tectonics, Earth's interior processes, surface landscape evolution, and climate, are coming more clearly into focus. By piecing together these data, it will be possible to reconstruct a deep-time history of tectonics on Earth that: (1) quantifies plate boundaries where crust is created and destroyed, (2) elucidates tectonic deformation in the continents that formed mountain ranges and sedimentary basins; (3) determines the changing configuration of ocean basins; and (4) establishes connections between solid-Earth systems and surface environments. These objectives will be achieved by designing whole-Earth geodynamic models that incorporate realistic estimates of crust and mantle rock properties. The data-driven approach proposed by the DDE has the potential to identify when and where plates have been deformed over hundreds of millions of years (Fig. 3-8; see Müller et al., 2019).

Some key questions are:

1) What drives plate tectonic motion and surface deformation?
2) How do Earth's plates move and deform through deep-time?
3) What is the connection between surface deformation and mantle convection?

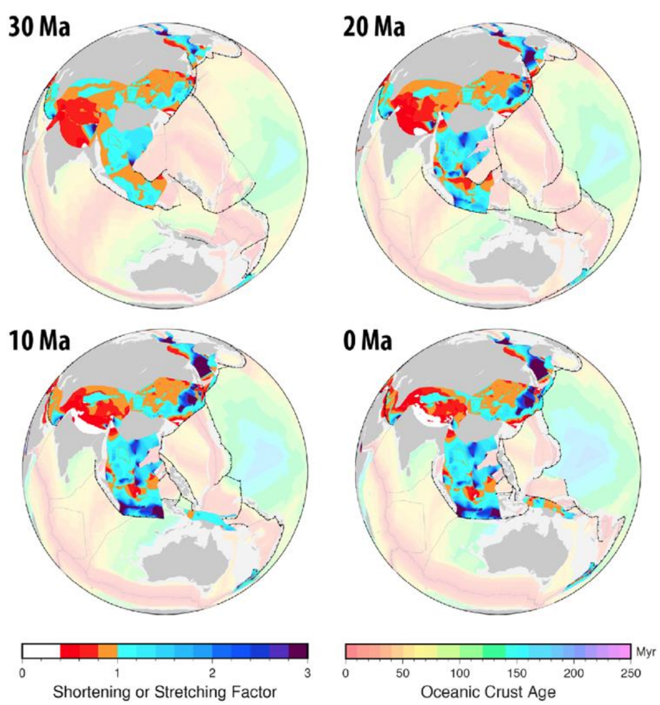

Fig. 3-8. Cenozoic plate deformation in Asia and Zealandia (Müller et al., 2019).

3.2.3 The Evolution of Life

Earth is the only known living planet in the universe. Understanding the origins and evolution of life on Earth continue to be among the greatest challenges in science. Insights are to be found in deep-time data on fossil organisms and their environments, as well as geochemical evidence that reveal biologically-induced changes in Earth's near-surface environment.

3.2.3.1 How did life and biodiversity evolve through time?

How did the modern biosphere evolve from the first primitive cells over 4 billion years of history? Fossil evidence reveals that life's evolution was replete with major biological innovations and transitions, mass extinctions and episodes when life recovered and flourished once more. A major DDE objective is to employ fossil data to unravel details of life's unfolding story.

"Biodiversity" is a measure of the number and variety of species on Earth. The fossil record suggests that Earth has experienced a long-term gradual diversification of life. However, it has proven extremely difficult to reconstruct Earth's biodiversity trajectory through deep time. Biodiversity information comes from two primary sources: phylogenies based on living organisms, and data gleaned from fossils, primarily macroscopic fossils from the last 600 million years. Information about biodiversity from the deep-time fossil record is fundamentally limited because only a small fraction of organisms living at any given time are thought to have been preserved as fossils. Many fossils, furthermore, have yet to be discovered. Consequently, a long-running controversy exists among paleontologists related to the construction of valid biodiversity curves – models that take into account both the preservation and sampling biases that create complexities in estimating the history of global biodiversity. Powerful new techniques, such as the analysis of paleontological data by application of artificial intelligence, show promise (Fig. 3-9).

Key remaining questions are:

1) How can a valid biodiversity curve be developed, taking account of inevitable sampling biases?
2) What resolution can we reach for a past diversity pattern?
3) What lessons do past changes in biodiversity hold for Earth today?

3.2.3.2 The chemical evolution of oceans and atmosphere

Understanding the chemical evolution of the oceans and atmosphere through geological time is crucial when considering Earth's ability to host complex life cycles, especially because marine and terrestrial ecosystems are interconnected. Understanding these changes also provides information relevant to the search for habitable environments and life on other planets.

Previously, insights into oceanic and atmospheric chemical evolution were derived primarily from measurements of geochemical signatures of elements that are sensitive to near-surface processes

– so called “paleo-proxies.” In the last two decades, a variety of new elemental and isotopic proxies have been discovered, significantly improving our knowledge of Earth’s chemical history. Recent studies have mostly focused on analyzing individual proxies – an approach that may result in significant biases in environmental interpretation. We suggest that multi-element and isotope approaches combined with geochemical modeling would improve our quantitative understanding of ocean chemistry changes through time (Fig. 3-10).

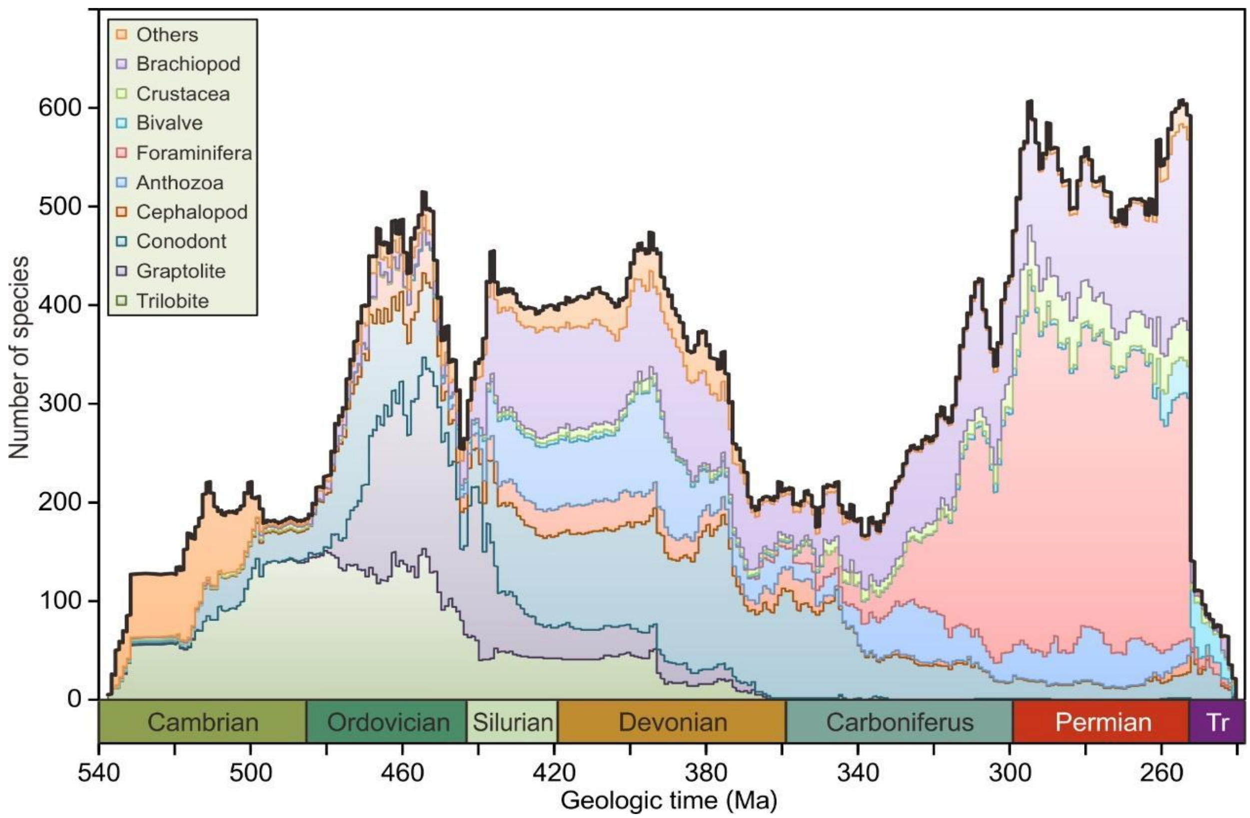

Fig. 3-9. Fan et al. (2020) employed paleontological big data and artificial intelligence algorithms to reveal marine biodiversity for 10 major fossil groups. This analysis demonstrates both dramatic mass extinction events and rapid recovery of biodiversity.

Our overarching goal is to understand how life and environments have co-evolved through geological time by integrating data from the geosciences and biosciences. To fulfill this “big-picture” goal, the DDE aspires to bring inorganic and organic geochemical analyses, models, experiments, and statistical data together to investigate the links between the biosphere, marine and atmospheric chemistry, and climate at variable spatial-temporal scales.

Some key questions are:

1) How can geochemical proxies that are sensitive to oceanic and atmospheric processes be used to reliably reconstruct ancient environments?

2) How do major fluctuations in near-surface chemistry relate to evolution of marine and terrestrial ecosystems?
3) What can our newfound understanding of ancient life and its environments reveal about modern ecosystems?

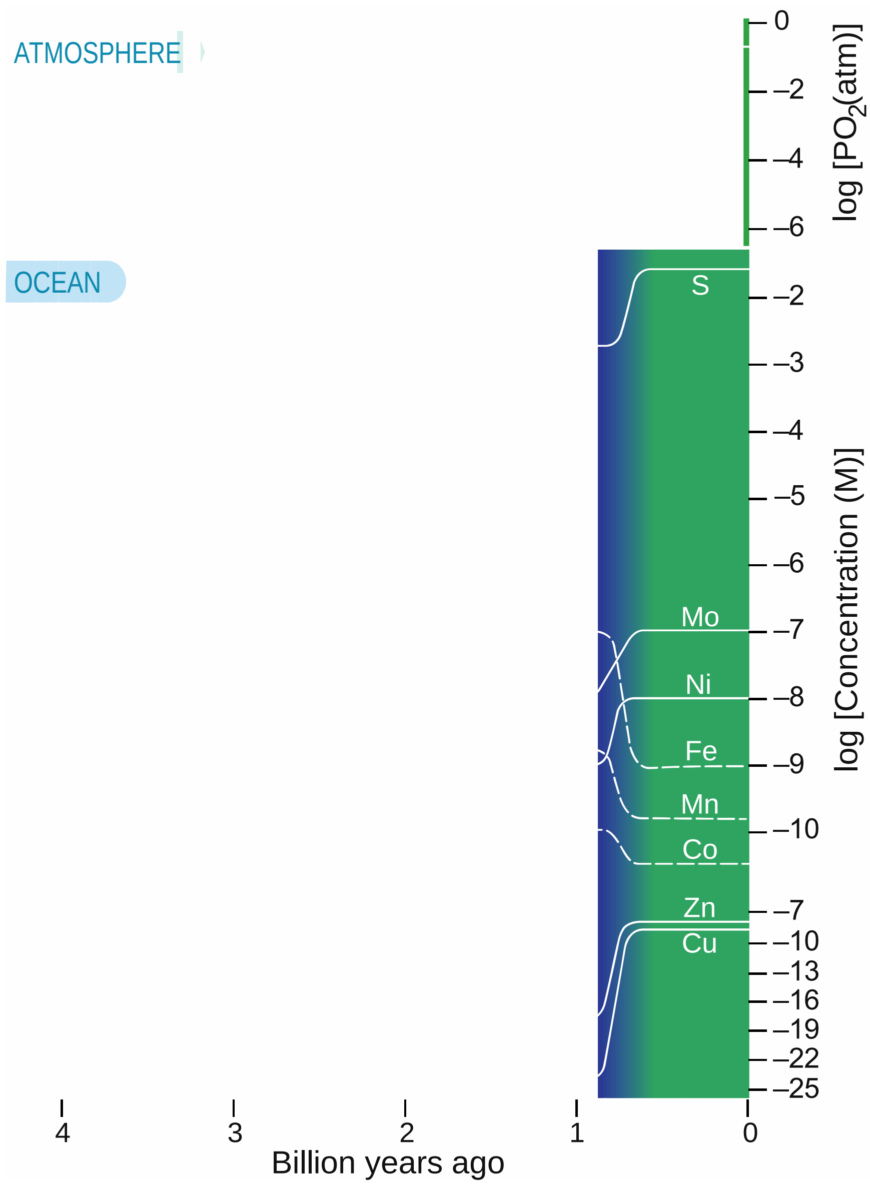

Fig.3-10. Anbar (2008) employed a variety of geochemical data to estimate changes in average ocean chemistry through deep time. Color gradations indicate a transition from anoxic, S-poor oceans before 2.4 billion years ago (light blue) to H2S-rich oceans between 1.8 billion and 800 million years ago (dark blue), subsequently giving way to complete ocean oxygenation (green). Different line styles are for clarity only; dashed lines are for elements with falling concentrations. These histories are approximate, based on simple geochemical models and inferences from ancient sediments. Ocean chemistry reflects complex interactions among many elements. Therefore, it is essential to consider multiple compositional variables simultaneously.

3.2.4 The Evolution of Climate

One of the most critical challenges for Earth scientists today is documenting the nature and extent of Earth's changeable climate – past, present, and future. Data-driven discovery offers a promising approach to enhanced understanding.

3.2.4.1 The co-evolution of atmospheric carbon dioxide and climate

The temperature at Earth's surface is determined primarily by the extent to which gases in the atmosphere absorb outgoing infrared radiation. The atmosphere is largely transparent to the Sun's incoming visible and ultraviolet radiation that warms the surface, but it is somewhat opaque to the infrared (heat) energy that radiates out into space. If it were not for the trapping of heat by the atmosphere, the average temperature of Earth's surface would be about -20°C. Thus, like a greenhouse, the atmosphere raises Earth's temperature from an inhospitable -20°C to its present more temperate temperature distribution. This natural temperature increase associated with atmospheric trapping of heat is the so-called greenhouse effect.

Carbon dioxide is an important greenhouse gas, whose atmospheric concentration has varied widely through Earth history. Therefore, the deep-time record of CO_2 in Earth's atmosphere is fundamental to understanding climate and how our planet evolved to become a habitable World. Evidence from varied sources – glacial deposits, geochemical signals, fossils, and more – reveal that over deep time Earth has experienced multiple extreme climate events, including several global-scale glaciation events, as well as periods of unusually hot climate. The most studied global glacial events are the Precambrian "Snowball Earth," the end-Ordovician glaciation, and the Late Paleozoic Ice Age. The most investigated hyperthermal events are the end-Permian/Early-Triassic "hothouse," the Cretaceous "greenhouse," and the Paleocene-Eocene Thermal Maximum. Those climatic events have dramatically shaped the evolutionary trajectories of Earth's biosphere, leading to mass extinctions followed by recoveries and adaptive radiations of new species.

Today, we are experiencing another interval of climate change, perhaps unmatched in the rate of temperature rise (Fig. 3-11). This warming episode appears to be closely linked to a human- driven rise in carbon dioxide as a consequence of burning coal and oil. Consequently, we are now living in an interglacial interval of the Cenozoic "icehouse" that has been perturbed by rapidly increasing CO_2 and temperature – changes that are leading to significant loss of biodiversity on land and in the oceans.

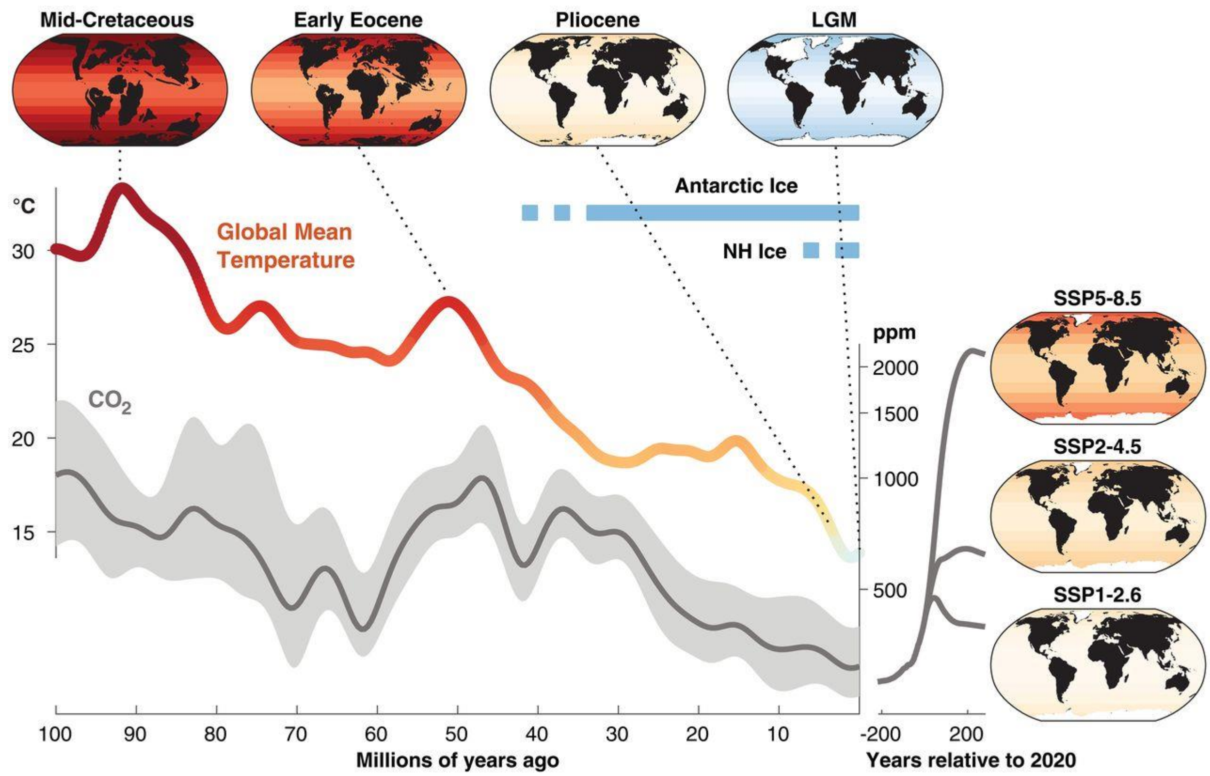

Fig.3-11 Paleoclimate context for future climate scenarios. Global mean surface temperature for the past 100 million years is estimated from benthic $\delta^{18}O$. Temperature colors are scaled relative to preindustrial conditions. The maps show simplified representations of surface temperature. Blue bars indicate when there are well-developed ice sheets (solid lines) and intermittent ice sheets (dashed lines) (Tierney et al., 2020).

Unravelling the tempo and mode of biosphere responses to extreme climate events have the potential to serve as a foundation for predicting current and future global changes. To do so, high-resolution robust reconstructions of Earth's deep-time climate and atmosphere history are needed. For the past several decades, geologists and geochemists have generated large amounts of data relevant to many important questions; as results have accumulated, researchers have increasingly gravitated towards larger compilations and statistical tools that can be implemented with Big Data studies on multiple geochemical proxies.

As the DDE moves forward, making these data comprehensive, temporally and spatially well-covered, and FAIR would contribute significantly to quantitative understandings of the roles of atmospheric change on climate. Important steps will include reconstructing levels of atmospheric carbon dioxide, oxygen, and other gases; estimating global mean and sea surface temperatures; documenting continental weathering and feedbacks (see below); and carbon cycling.

Key questions include:

1) When and how did CO_2 accumulate in the atmosphere and in the oceans?

2) How and to what extent have Earth's climatic/environmental perturbations shaped the evolutionary trajectories of the biosphere?
3) What caused and terminated extreme climate events?
4) How and why has Earth's climate stayed habitable continuously over the past 500 Ma?
5) What past climate changes inform our understanding of future global changes, and can we identify ways to minimize human impacts on the atmosphere, ecosystems, and biodiversity?

3.2.4.2 The role of continental weathering in moderating climate

Continental weathering has long been a bridge that connects Earth's atmosphere, hydrosphere, soils, and biosphere. Over geological timescales, continental weathering has played a major role in maintaining Earth's habitability by regulating atmospheric carbon dioxide (CO_2) levels by removing CO_2 in high-CO_2 atmospheres, while accumulating CO_2 during periods of relatively low atmospheric CO_2. Rock weathering releases and delivers nutrients such as P, Si, and Fe, and bio-essential trace metals, supporting the primary productivity of the biosphere and enhanced organic carbon burial, which further gave rise to atmospheric O_2. In addition, weathering processes lead to the formation of numerous important economic resources, such as aluminum-rich bauxite deposits.

The negative feedback between atmospheric CO_2 concentration and weathering CO_2 consumption is one of our planet's the most important feedback mechanisms, because it reduces the chances of out-of-control greenhouse or icehouse conditions in response to perturbation of the carbon cycle. This close link of weathering to Earth's climate makes the record of weathering potentially a key component of DDE. However, the record of weathering is complicated by the highly spatial heterogeneity of weathered terrains (e.g., mountains, deserts), the diversity of paleo-weathering signals and proxies (minerals, sediments), and contributing substrates (detrital sediments, igneous rocks and minerals). Application of Big Data and machine learning may offer a strategy for synthesizing these complicated records of weathering. Additionally, numerical climate models that assimilate these proxy data will need to be developed to understand how weathering has evolved over time (Boucot et al., 2013).

Some key questions are:

1) How does weathering respond to major tectonic and biological changes?
2) What is the contribution of weathering to the past changes of environment and biosphere?
3) How do feedbacks associated with weathering help to maintain the long-term habitability of the Earth?

4. A Roadmap for Deep-Time, Data-Driven Discovery

Scientific discoveries related to Earth's evolving materials, geography, life, and climate are relevant to some of the most profound unanswered questions about the natural world, while bearing a close connection to critical societal needs. A central motivation of the DDE Program is to apply discoveries of Big Data geoscience to address these scientific questions and societal concerns.

Effective strategies for successful deep-time, data-driven discovery depend on harnessing the power of data analysis and visualization to accelerate discovery. To this end, the DDE is developing a Roadmap for Discovery, based on three elements: (1) constructing the Geoscience Knowledge Graph; (2) accumulating Geoscience Data Resources; (3) developing a Platform for data analysis and visualization. These three elements, collectively, will provide an open-access data Smorgasbord for geoscientists – with the DDE infrastructure, every flavor of Earth science data can be selected, integrated, and analyzed. DDE aims to provide an open platform for linking existing deep-time Earth data and integrating geological data that users can interrogate by specifying time, space and subject (i.e. a 'Geological Google') and for processing data for knowledge discovery using a knowledge engine (Deep-Time Earth Engine) that provides computing power, models, methods and algorithms (Fig. 4-1). DDE can help scientists with time-consuming data cleansing and processing so they can focus on research topics and discoveries to address the scientific questions and societal needs.

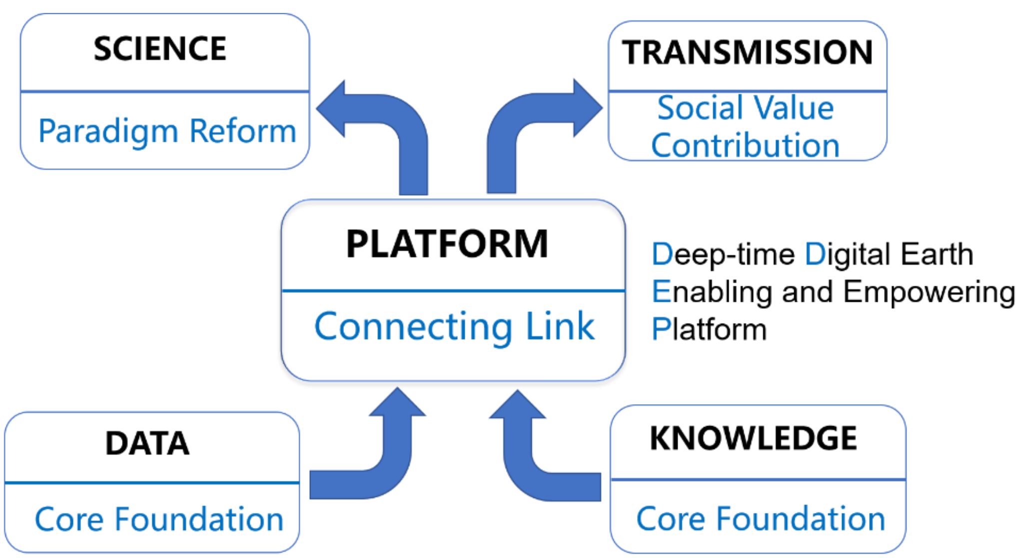

Fig. 4-1. Concept of DDE system.

4.1 The Geoscience Knowledge Graph: A Key to Innovation and Applications

Human society is evolving at an ever-accelerating pace and with that acceleration has come an unprecedented increase in the amount of data in many public, private, and corporate domains (Ansari et al., 2019). These massive data resources bring new challenges to storage and retrieval technology.

A particular challenge is presented by geoscience data, which occur in exceptionally diverse and complex formats and types, from photographs of rock thin sections, to geological maps, to seismic signals, to measurements of ecosystems in four dimensions. The resulting rich semantic relationships make the construction of Earth data and knowledge extremely cumbersome. We lack a suitable geoscience ontology model that encompasses the many interconnected facets of Earth systems. Therefore, it is imperative to construct a knowledge graph for geoscience. In this endeavor, the semantic heterogeneity of geoscience terms and the variability of data collection, evaluation, compilation, and retrieval standards are major problems to be solved. Ultimately, the construction of an interoperable, semantically-linked, large-scale, and heterogeneous geoscience database will provide an unprecedented opportunity for data mining, knowledge extraction, and reasoning.

A knowledge graph (Box 3), also known as a semantic network, represents a network of real-world entities—i.e. objects, events, situations, or concepts—and illustrates the relationship between them. This information is usually stored in a graph database and visualized as a graph structure, prompting the term knowledge "graph". A knowledge graph is made up of three main components: nodes, edges, and labels. Any object, place, or person can be a node. An edge defines the relationship between the nodes. As far as the current computer technology is concerned, a knowledge graph is one of the more effective ways to express such complex relationships . The DDE knowledge graph will consist of four application layers: the DDE Knowledge Editor, Crawler, Provider, and Enabler. In the future, we will explore major scientific issues in the field of geosciences based on this knowledge graph approach, which will facilitate new approaches to research in the geosciences.

4.1.1 Structure of the DDE Knowledge Graph

The DDE knowledge graph contains four application layers: DDE Knowledge Editor, Crawler, Provider, and Enabler.

The *DDE Knowledge Editor* is a collaborative tool to create a knowledge structure, by which scientists can transform data items into standard and regular expressions. They create data objects using a Consensus Mechanism and Motivation Mechanism. The Knowledge Editor includes a variety of intelligence tools, such as Cloud Platform and Visual Dashboard. Users extract and merge data via this system. In this process of data resource development, we emphasize the importance of data security, including intellectual property protection, data/knowledge privacy, and encryption technology.

The *DDE Knowledge Crawler* is to construct a comprehensive geoscience knowledge graph by using modern information technology from massive structured and unstructured geoscience literature, including journals, books, governmental reports, corporate records, and more, which contain a huge amount of geoscience knowledge.

The *DDE Knowledge Provider* is an artificial intelligence resource that serves global users and plays the role of education, training, and guidance in geoscience research. It mainly consists of four functional modules: knowledge display, intelligent retrieval, knowledge repository, and automatic recommendation.

- Knowledge Display includes diversified knowledge display methods, including a dictionary, tree, graph style, and other methods.
- Intelligent Retrieval supports the retrieval of text, image, video, and other information in a variety of ways. It also has the option of sorting outputs according to relevance and importance.
- Knowledge Treasure is a feature that allows each user to build a unique knowledge "treasure house" to establish a new data resource or to summarize the existing data related to one or more facets of DDE Knowledge.
- Automatic Recommendation helps to identify users' interests and to provide relevant data of interest.

The *DDE Knowledge Provider* also provides users with standard specifications and unified interfaces to call knowledge graph in other applications.

The *DDE Knowledge Enabler* provides advanced algorithms, models, and tools, which helps promote and accelerate the development of new applications to realize the application and implementation of DDE data, and make scientific discovery.

Knowledge Provider and Knowledge Enabler are both open platforms.

4.1.2 Method of Constructing the DDE Knowledge

DDE knowledge is constructed by employing the Geoscience Knowledge Representation Model, which consists of nodes representing concepts (entities) and directed edges between pairs of nodes representing relationships. However, geoscience is a heterogeneous science that studies the formation, evolution, and interaction of very different spheres of Earth, including the atmosphere, oceans, geography, geology, geophysics, paleobiology, and more. Such diversity demands similarly complex and diverse disciplinary knowledge systems. Consequently, problems such as low efficiency, inaccuracy, and complex reasoning in the extension of spatiotemporal characteristics and complex relations invariably arise. Considering the characteristics of geoscience knowledge, such as complex relationships among systems at widely varying scales of space and time, it is challenging to adapt existing graph models to represent and design a geoscience knowledge representation that integrates all relevant characteristics. Development a Knowledge Representation Model is an important DDE objective.

Knowledge representation is the basis of constructing the computer-compatible and calculable knowledge graph, and it is also an important step in the process of knowledge communication. Existing geoscience knowledge representation models generally represent the temporal, spatial, and attribute information, which is insufficient to model complex, multi-domain geoscience knowledge (Zhang et al., 2020). Consequently, the fundamental challenge is how to construct a representation model of a comprehensive geoscience knowledge graph across all spatiotemporal dimensions. To tackle this complex challenge, DDE proposes a basic graphical model of the self-adaptive representation of all-domain geoscience knowledge (Fig. 4-2), which is composed of a complex spatiotemporal information representation model (entity object representation model) and geo-entity object relationships (edges).

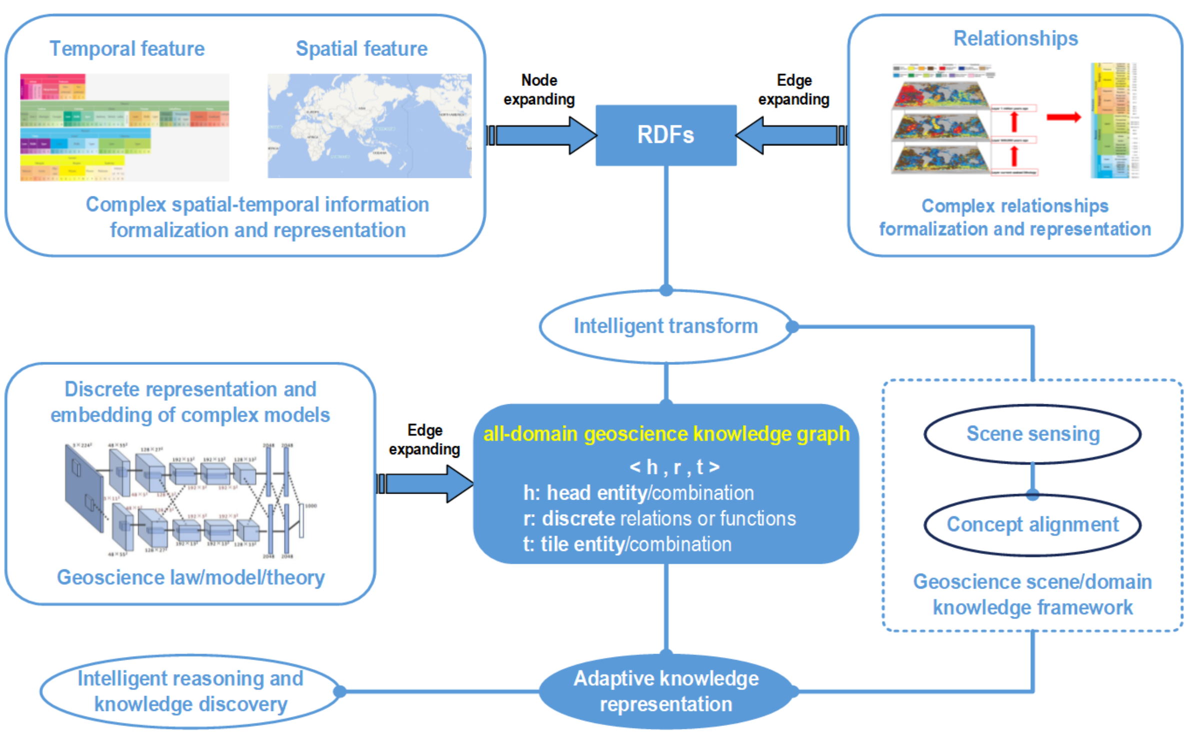

Fig. 4-2. Self-adaptive representation model of all-domain Geo-Knowledge Graph (Zhou et al., 2021)

Generally, strategies of constructing knowledge graphs are divided into "bottom-up" and "top-down" approaches (Meilicke et al., 2019; Zhao et al., 2018). The construction of the DDE geoscience knowledge graph will combine these two methods. In the "top-down" stage, DDE proposes the construction through crowd intelligence collaboration. Geoscience knowledge encompasses a wide range of data, most of which comes from geoscientists themselves. To manually input the knowledge and experience of these experts with traditional methods requires a high degree of time-intensive collaboration. With the rapid development and distribution of intelligent mobile devices that can collect and report geospatially-constrained data, the option of crowd-sourced data acquisition provides a new model to gather global data.

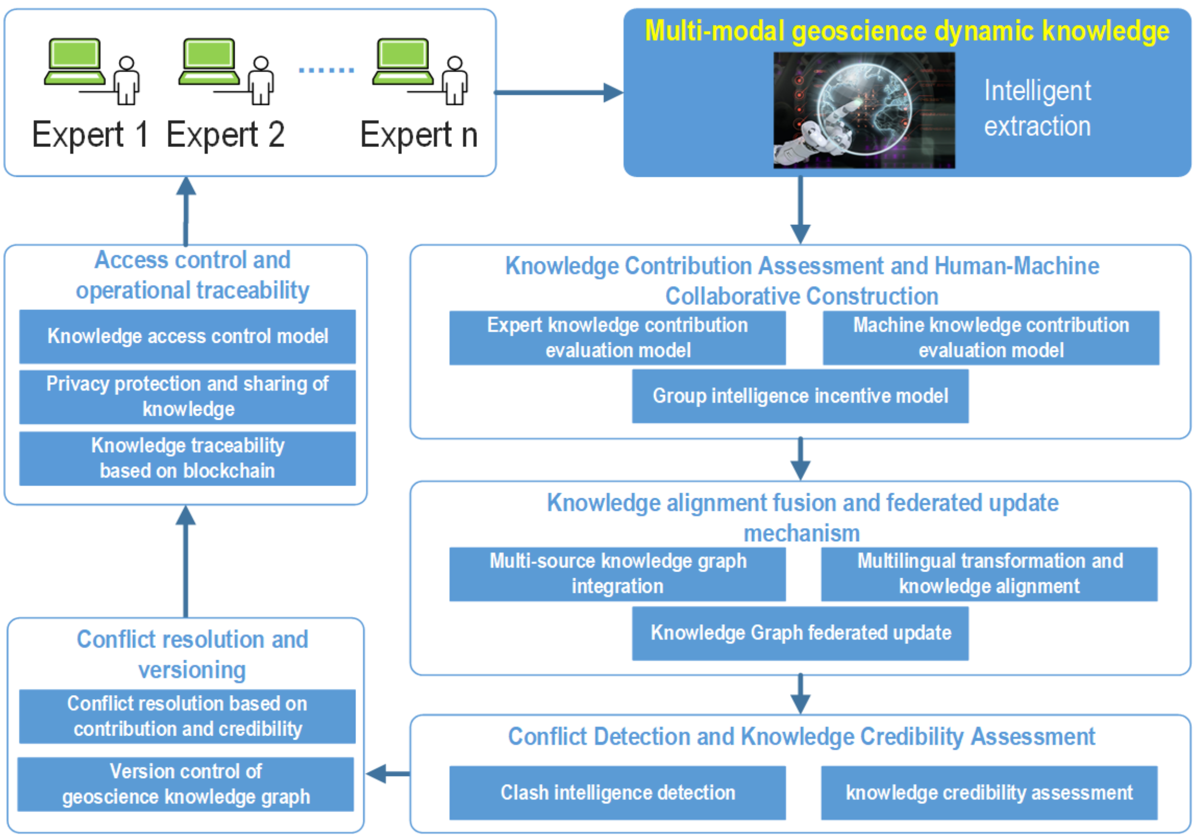

Fig. 4-3. Construction Framework of Geoscience Knowledge Graph through Federated Crowd Intelligence Collaboration (Zhou et al., 2021).

By relying on the observations collected by numerous individuals around the world, the DDE geoscience knowledge graph can be updated in real time. Therefore, DDE proposes a federated crowd intelligence knowledge graph construction framework with expert knowledge as the core (Fig. 4-3). In this endeavor, several key technical issues need to be resolved:

- Contradictions and conflicts will inevitably arise as different experts evaluate and analyze the large-scale geoscience knowledge graph constructed through crowd intelligence collaboration.

- Planning and establishment of a sustainable model for gathering and evaluating crowd-sourced data is critical for construction of the global geoscience knowledge graph.
- Large-scale collaboration of geoscientists from around the world requires a solution to the problem of multilingual representation of knowledge.
- It is essential to establish consistency-based methods of creating a multi-source knowledge graph based on rules, statistics, and deep learning.

Massive structured and unstructured geoscience literature, including journals, books, governmental reports, and corporate records contain a huge amount of geoscience knowledge. Text mining represents an important method to construct a comprehensive geoscience knowledge graph by using modern information technology to extract and update geoscience knowledge. In view of this largely untapped text mining potential, we propose development of methods for dynamic, automated knowledge graph construction (Fig. 4-4).

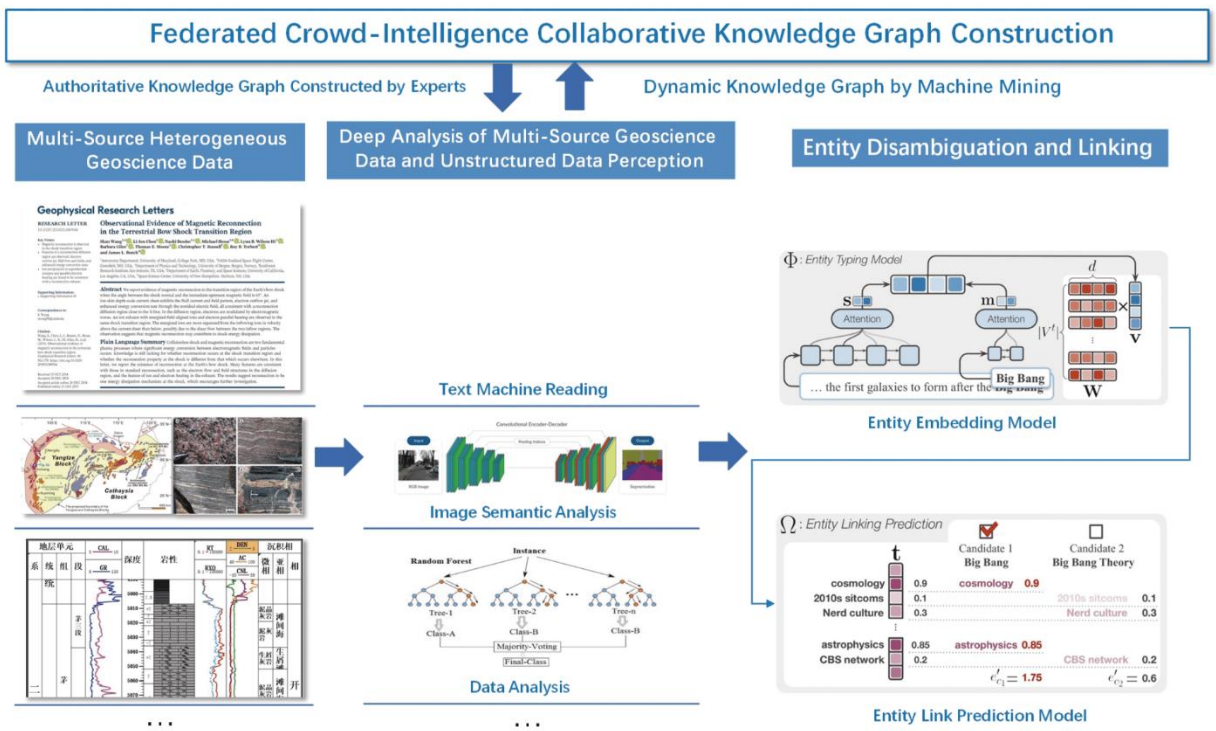

Fig. 4-4. Construction Framework of Dynamic Knowledge Graphs (Zhou et al., 2021)

Three steps are required to achieve effective text mining. First, to realize text association and multi-source data perception, massive unstructured text materials must be categorized, while the associated attributes of data from the same source must be marked. On the basis of marking, the text-mining algorithm segments the processed text, extracts plain text, and analyzes syntax. Graphs, texts, and numbers from different sources with certain similarities are marked and associated by adopting text-matching and statistical learning methods. The geoscience literature often contains maps, tables, and other professional components; therefore, text-mining methods must be trained to understand the meanings of map symbols and identify various kinds of spatial relationships.

The second step in implementing text mining is entity object and knowledge extraction based on keywords. One of the keys of deep learning is developing and applying high-quality training sets. It is crucial to develop efficient and credible unsupervised learning algorithms, such as object extraction based on keyword graph pattern. The algorithm first quantifies statistical characteristics of geo-entities based on word segmentation results of massive texts, coupled with the Term Frequency and Inverse Document Frequency (TF-IDF) algorithm. Building the list of common words according to the sorting results, we will construct the language network graph of massive texts. Based on matching and sorting sizes and vectors of graph features, we will search for important words or phrases in the language network graphs. Ultimately, the resulting text pattern recognition will facilitate geoscience knowledge extraction based on unstructured texts.

The third and final step in text mining is the disambiguation of knowledge and the construction of dynamic geoscience knowledge graphs. To solve the ambiguity and conflicts of geoscience knowledge from multiple data sources, the problems of polysemy and synonymy are classified through deep reinforcement learning that is characterized by the specific spatiotemporal semantic association of entity concepts. Further, the method aligns attributes to eliminate knowledge conflicts by training the information source through feature learning, which is characterized by source attributes.

At the heart of DDE efforts in knowledge construction is knowledge reasoning, which refers to establishing new relationships between entities, understanding the characteristics of evolution of knowledge systems, and discovering new knowledge through computer reasoning and based on the relationships among entity concepts in the knowledge graph. At present, the common methods for knowledge reasoning include symbolic reasoning and statistical reasoning. The core of symbolic reasoning is to use rule of relevance to infer new entity relationships from existing ones and detect possible logical conflicts within them. Alternatively, statistical reasoning employs machine learning and other methods to develop new relationships among entities in the knowledge graph based on past experience and analysis, and verify or infer assumptions using Maximum-a-posteriori (MAP) and other statistical methods.

4.1.3 Applications of DDE Knowledge

The applications of DDE Knowledge can promote the overlap and integration of geoscience, information science, and data science as well as accelerate the development of these disciplines. Through the combination and derivation of Editor, Enabler, DDE Knowledge can facilitate many typical applications. Three typical applications illustrate DDE activities after the initial construction of DDE Knowledge.

4.1.3.1 Construction of a High-resolution Geological Time Scale based on DDE Knowledge

Development of a unified spatiotemporal framework is a prerequisite and cornerstone for geoscience research, and thus it is a DDE ambition to establish a unified deep-time framework with ten-thousand-year time resolutions. With the development and expansion of the Geoscience Knowledge Graph, the realization of a high-resolution geological time scale becomes possible.

Two developments are necessary to achieve this DDE goal. First, we must develop intelligent methods for automatically extracting temporally-constrained basic concepts, terms, specifications, stratigraphy, paleontology, geochronology, astronomical cycles, and isotope chronology. We need to formulate coding specifications suitable for describing Geoscience Knowledge Graphs based on international standards and specifications, and construct a machine-understandable ontology of Knowledge Graphs for stratigraphy, paleontology, geochronology, astronomical cycles, and isotope chemostratigraphy.

Second, we have to develop algorithms based on high-performance computing and artificial intelligence technology to compare and calibrate time and rate of major biological environmental events in the context of the geological time scale. We will adopt an iterative approach, refining the geological time scale of different geological periods step by step, such as the new generation of geological time scale from a million-year to thousand-year level in the Paleozoic Era. Ultimately, we will create a geological time scale spanning the entire Earth history.

4.1.3.2 Geoscience Knowledge Editor

The DDE Knowledge will promote extensive participation of the Earth science community. Specific interest groups will distill the existing domain ontologies for curating geological knowledge of deep-time Earth. Scientists will collect and organize the knowledge and display it in a machine-readable structure to produce their knowledge systems. The information for geological entities (e.g. lithostratigraphic units) in the Knowledge Graph will include official names, synonyms, definitions, and relationships to other entities, as well as their historical meanings. DDE has developed a cloud-based platform for analyzing and building knowledge systems. The platform has functions such as knowledge entering, reviewing, browsing, querying, online updating, intelligent interacting, and sharing.

4.1.3.3 Intelligent Geological Map Editing and Mapping

Intelligent geological mapping based on DDE Knowledge is driven by geo-knowledge, cartographic knowledge, and intelligent selection of big data in the context of cartographic conventions. This effort implements autonomous judgment in conjunction with collocation of various mapping resources, including mapping data, models, templates, and methods. Thus, geological mapping aided by machine learning can promote the development of automatic and intelligent cartography. Two core aspects of intelligent cartography include the external task of knowledge-driven workflow, coupled with the internal task of map-making assisted by data processing, overall design, content organization, and symbolic representation.

4.2 Harmonizing Earth Science Data

With the continuing development of new observational and experimental methods, geoscience data are growing at an unprecedented rate. Therefore, it is necessary to consider how to help geoscientists discover and access these data quickly, accurately, and concisely by linking existing databases, collecting unstructured or even non-digitized dark data as well. The development of such data resources is the foundation of data-driven geoscience research (Hazen et al., 2019; Stephenson et al., 2020).

The mission of DDE's Big Data Program (DDE Data) is to harmonize and share data on Earth's evolution. To this end, one of the core goals of DDE Data is to link and integrate multi-source, heterogeneous geoscience data (Wang et al., 2021). Therefore, the main DDE Data tasks include: (1) formulation of an open and shared global scientific data resource system; (2) the design of standards and norms for data integration and sharing of the Earth evolution science; and (3) the construction of a DDE Data infrastructure and its corresponding software platform system.

4.2.1 Integrating FAIR data to form connected data hubs

Understanding the long-term evolution of Earth systems and their controlling factors, including anthropogenic contributions, relies to a large extent on developing new methods for integrating and querying different types of Earth observations and models of geoscience data. In the new era of data-driven Earth systems, innovative platforms and programs need to be developed for efficient utilization of, and deep learning from, geoscience data. The DDE Data aims to provide interoperability of databases created and maintained by national geological surveys, government agencies, academic organizations, and industry, as well as databases developed by individual scientists, teams of scientists, or consortia through individual scientific programs and projects. DDE Data will address these critical issues by developing and delivering community-based solutions. The first step is to promote the development and adoption of database specifications for assembly of distributed systems of connected geoscience informatics populated with FAIR data (Fig. 4-5).

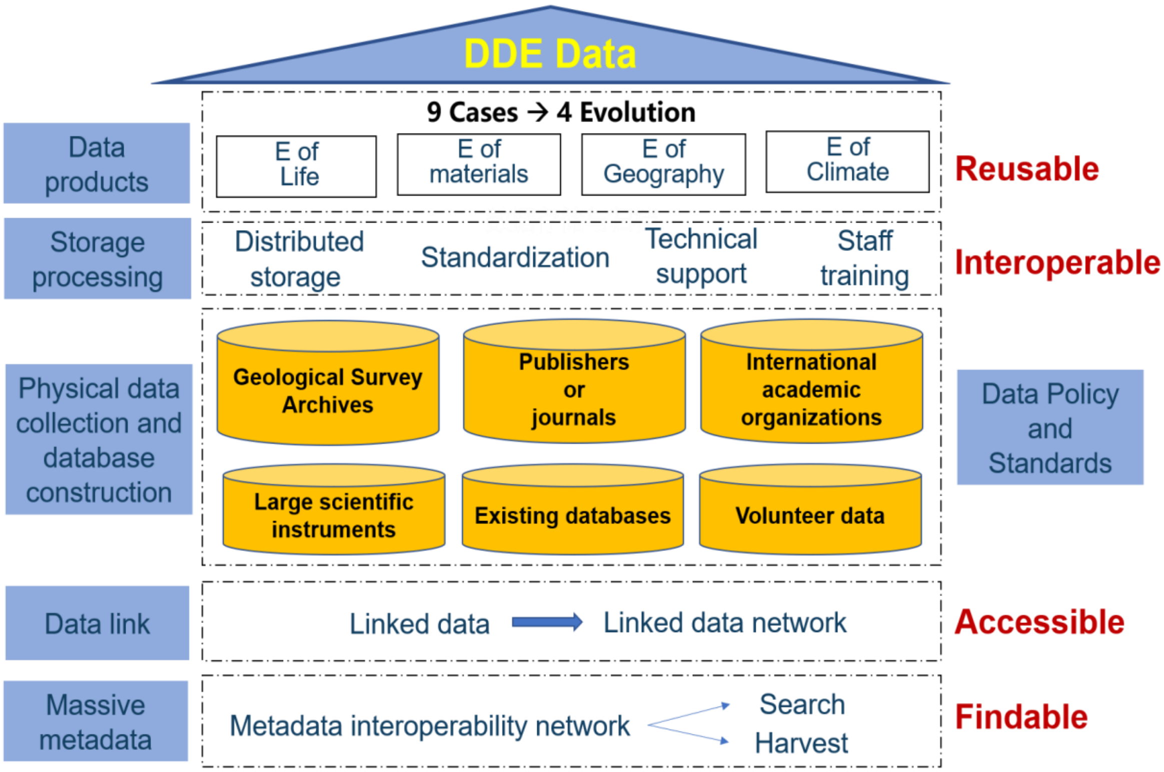

Fig. 4-5. DDE Big Data System Architecture

For promoting integration and open sharing of global Earth evolution science data, it is necessary to clarify the interests and demands of participants in open data sharing, establish effective incentive mechanisms, benefit distribution mechanisms, evaluation mechanisms, and diversified open sharing strategies and mechanisms for achieving international cooperation and orderly advancement of the deep earth science data open sharing (Mokrane et al., 2016). The main contents include: DDE Data intellectual property policies, data exchange mechanisms, data sharing alliance mechanisms, data publishing mechanisms, volunteer data sharing mechanisms, data use policies, member management methods, and data security protection mechanisms (Lin et al., 2020).

A key is to establish the DEE Data standard specification environment, and to construct the data standard system covering the whole life cycle process of data access, identification, processing, integration, and processing, service. DDE Data must establish technical standards in accordance with international and domestic compatibility principles and solve the problem of DDE big data integration. Specifically, it includes: (1) establishing reference models and systems of DDE data standards, which are divided into three layers: guiding standard, general standard, and special standard; (2) establishing the classification and coding system of DDE data, and developing a catalog system that can be used for data exchange and users' convenient access to data; (3) studying the basic data unit supporting studies of Earth evolution, forming DEE data integration standards, and establishing the granularity of the basic unit for the long-term continuous integration and sharing of deep-time earth evolution data.

The management and application of data identification includes identification registration and distribution management, metadata query and statistics, and register data intelligent analysis. Based on the unified metadata standard and data interoperation service specifications, the system will achieve compatibility with existing international, national, and industrial data resources. DDE Data will support the distribution and rapid integration of multi-source, heterogeneous deep-time earth science data, as well as indexing and managing data, implementing efficient data queries, browsing, and extraction, to offer a one-stop service for Earth science data.

4.2.2 Building various data nodes to facilitate data sharing

The DDE Data has been developed and constructed to enhance the integration and sharing of deep-time data for geosciences. The system is deployed and operated according to four types of DDE Data nodes (national nodes, academic nodes, industry nodes, and community nodes). Under DDE's unified metadata standards and directory specifications, the integration, exchange, and sharing of global geoscience data will be realized. Using DDE data sharing mechanisms and standards, deep-time geoscience data from various sources will be integrated, including national node data, professional data of academic organizations, data of international big science programs, data of publishers' literature, unpublished data from scientists and volunteers, and other potential data sources. Through the implementation of DDE Data Alliance, to promote the cooperation of publishing groups in the field of geosciences at home and abroad, and to realize data exchange, integration, and storage of published data in a uniform DDE consortium. Through principles of network refunding and data sharing, the effective integration of scientist group data and volunteer data can be realized (Nicol et al., 2013).

Take an example, in the construction of national node, the DDE-China Construction will be carried out first and serve as a reference model (Fig. 4-6). The main construction contents include classification analysis of national node construction, sharing and update strategy of the national node, construction of the national node professional data system, setting and registration of data in various subject, and rights of national node and user system.

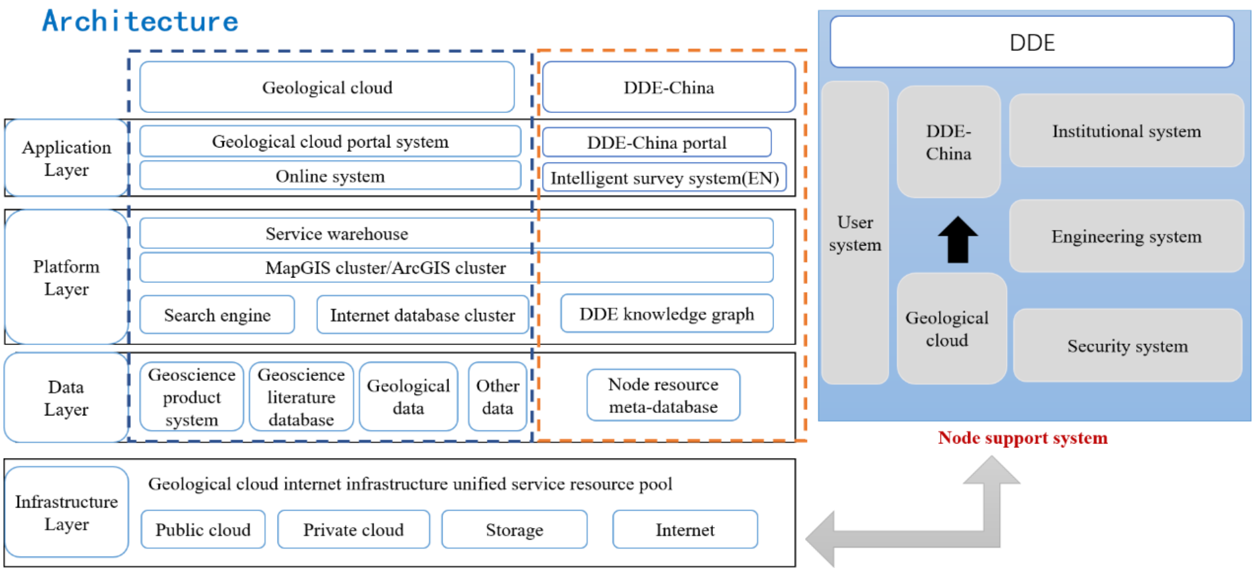

Fig. 4-6. DDE China Portal Architecture

DDE aspires to cover the full spectrum of solid Earth. A challenge is the diversity of data types and formats from different disciplines, each of which has its own data sources, characteristics, structures, and standards. Therefore, DDE Data takes into account disciplines, regions, data source channels, and other dimensions, and is implemented step by step to promote the construction of a full series of subject databases, including stratigraphy, paleontology, sedimentology, palaeogeography, geotectonics, paleomagnetism, mineralogy, magmatic petrology, metamorphic petrology, geochronology, mathematical geology, geomorphology, and other subject databases.

4.2.3 Linking knowledge systems and artificial intelligence to data discovery

One of DDE primary goals is to facilitate machine learning (ML) and artificial intelligence (AI) to support data-driven discovery in the Earth sciences. ML and AI have great potential to contribute to the acquisition of massive and heterogeneous data and automation of data discovery flow processes. Both are dedicated to studying how to use experience to improve the performance of the system itself through computing. In short, the main purpose of both is to find patterns hidden in data (Provost and Kohavi, 1998a, 1998b). ML and AI have been used in geoscience for several decades for various purposes ranging from prediction and simulation to multivariate analysis. The main challenge for facilitating greater use of ML and AI in geoscience is harmonizing and standardizing geoscientific data and developing specialized ML and AI for complex geoscience problem solving. DDE Data aims to coordinate and support these processes by providing a connected and open platform to provide “one stop” processing and analysis, thus overcoming this challenge.

Earth scientists are still facing many profound scientific issues, which can only be addressed through international collaboration. Earth has undergone more than 4.5 billion years of history, involving several major evolutionary stages including the origins of life and plate tectonics. However, due to the long history of acquiring deep geological records and the limitations of techniques to reconstruct that record, there has been a lack of adequate theory and methods for quantitatively describing Earth's spatial and temporal co-evolution.

Big Earth data show that major geological events such as the formation of supercontinents, the mass extinctions, atmospheric oxidation events, and episodic formation of mineral deposits are consistent with subduction, plume, and large-scale magma activity (Cheng, 2019). How these simultaneous internal processes control the occurrences of extreme events in the lithosphere or in the crust has long been a focus for geoscientists. These types of profound geological problems involving interaction of multiple earth systems over a wide range of spatial and temporal scales require complex analysis and attention to details emerging from many disciplines.

The main purpose of acquiring and using scientific data is to solve problems. The amount of information and knowledge that one can extract from data is therefore the primary measure of success of data utilization. The integration of human ideas and machine learning processes can be augmented and automated by modern AI technology and various advanced semantic knowledge engines. Information technology and big earth data availability are the key elements of the knowledge engine and models. These flows can be extended to combine machine learning processes, information management, and infrastructural organizations. Researchers can develop their own models for specific tasks using the DDE knowledge engine supported by DDE Data. The models can be edited, modified, and used to solve similar problems and, in turn, to refine the models through positive feedback processes. Lots of science discovery chances will be emerged in these interactive processes.

4.3 Data-driven Collaborative Cloud Computing Platform for Geoscience

Going through empirical paradigm and system paradigm, in the era of big data, the research scope of geology is transforming from single discipline to multi-disciplinary earth system science, a new data-driven scientific research paradigm is arising.

The DDE platform responds to the new research paradigm of "data-model-knowledge" as the core, multidisciplinary collaboration and mutual penetration. It is a spatiotemporal system-based four-dimension space-time coupled digital Earth operation platform, whose emphases are on deep-time, cloud, data, computation, and collaborative exploration compared with traditional geoscience platforms. Therefore, the DDE Platform's features are mainly:

1) Possess the abilities of joint discovery and interactive exploration analysis;
2) Construct robust cloud-native infrastructure, supporting Platform's resources allocation and direct applications build-ups;
3) Provide visualization and interoperability of multi-time-scale for geoscience scenarios, events, and features;
4) Be capable of time-labeled data management, distribution, and service;
5) Equipped with abilities to arrange abundant complex models, perform real-time computing, and speed up research.

The DDE Platform aggregates advanced cloud-native technologies, abundant data resources, multidisciplinary research models, and scalable computational power. Relying on big data, cloud and AI, DDE Platform facilitates global geoscientists to collaboratively discover and innovate efficiently and intelligently, which would play a significant role in expanding the boundary of human-being civilization.

4.3.1 Structure of the DDE Platform

The DDE Platform is built upon DeepNet to form a cloud native infrastructure basement (Fig. 4-7), fueling each engines with ample resources, e.g. ACK container service, ECS elastic server, relational database, and object storage, etc. The DeepNet is a hybrid cloud with global network services, massive storage capacity, and high-performance computing capabilities：

- *Network*: To serve geoscientists around the World, the DDE Platform must be accessible anywhere and anytime, which can be achieved through globally-distributed data centers provided by the universal hybrid cloud platform. Moreover, DDE's high-speed content distribution network (CDN) enables global users to obtain the required content locally, reducing network congestion, increasing response speed, and ensuring system access.
- *Storage*: The DeepNet will provide almost infinite data storage resources for managing large publicly available Earth science datasets as well as private datasets uploaded by individual scientists and organizations. As the Earth science data resources are usually multi-source, heterogeneousa with large amount and single storage system cannot fully meet the requirements, the DDE will provide polymorphic hybrid data storage and management resources on the cloud, such as relational databases, distributed databases, and distributed file systems.
- *Computation*: The DDE Platform requires a variety of scalable computation abilities. A cloud computing system is essential for analyzing data, conducting experiments, and sharing results, as it is very time-consuming to download massive Earth science datasets to local computers. Moreover, advanced high-performance computing ability, such as message passing interface (MPI), is also required because some specific models contain a large amount of iterative calculation and require fast communication between in-memory processes. In addition, the

DDE Platform also has built-in support of distributed deep learning systems with different back-ends, such as TensorFlow and Pytorch.

- *Security*: The DDE infrastructure must ensure the technical security of its server, network, data, and applications. Generally, the *universal cloud platform* provides comprehensive security protection mechanisms, such as data backup, vulnerability management, and intrusion detection. This security protection mechanism should be invisible to scientist users so that they can focus only on data analysis and Earth science problems.

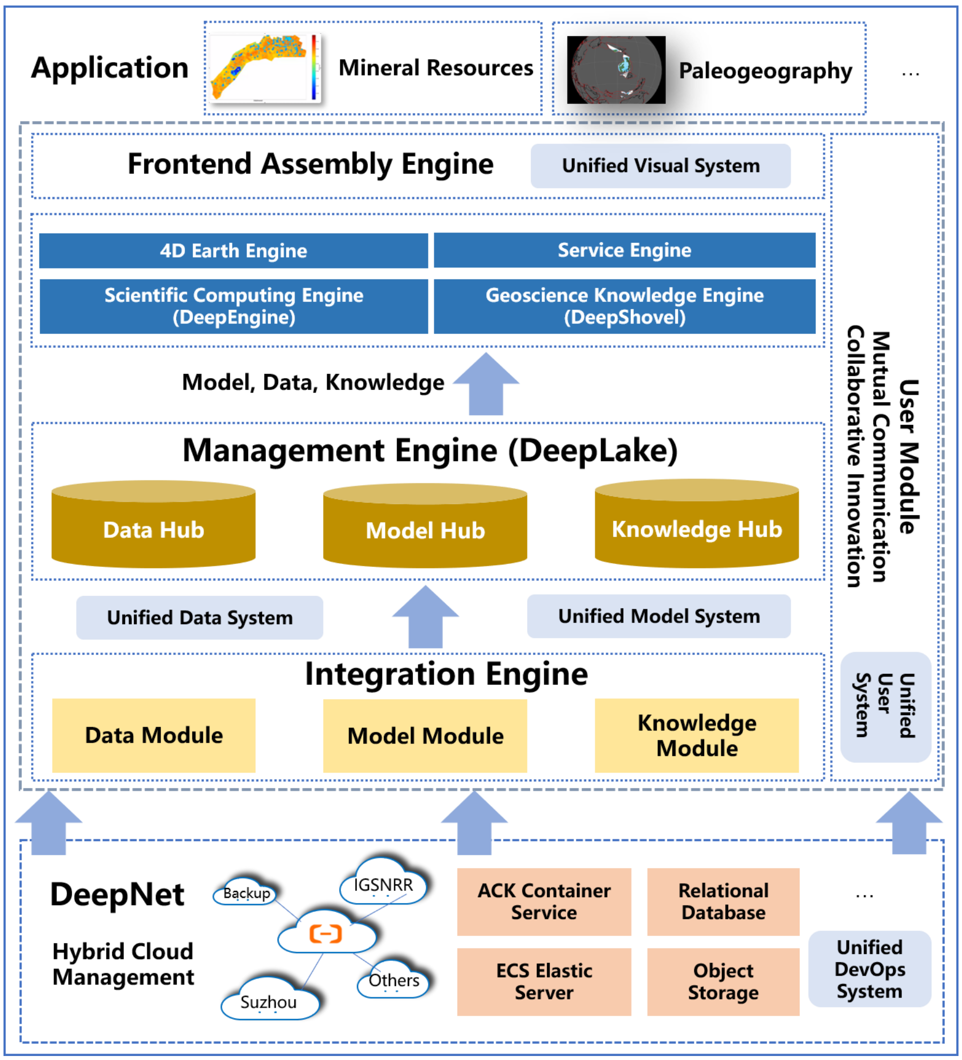

Fig. 4-7. Structure of the DDE Platform

Based on the unified user system, DDE Platform creates unified user modules to realize "once login, multi-station sharing" capability for all DDE website groups as the basics for collaborative communication, cooperation, and innovation.

The DDE Platform integrates the products of the DDE Knowledge group, DDE Data group through engines and integrates the unified data system and the unified model system to import the data hub, model hub, and knowledge hub in DeepLake. In addition, DeepLake provides resource scheduling services, including the scientific computing engine, the geoscience knowledge engine, the four-dimensional earth visualization engine, and the business-oriented engine.

In general, the DDE Platform is a one-stop platform for geoscience data preprocessing and analysis, which allows researchers to access and analyze existing data using a wide range of algorithms and models to tackle significant questions of Earth sciences, such as evolution of materials, geography, life, and climate. It is also a kind of global e-science platform, which shares scientific outcomes on the internet, such as publications, code, and data. It aims to provide an integrated scientific research environment for the CPU-intensive or IO-intensive data analysis experiments of the Earth sciences conducted collaboratively on a globally distributed IT infrastructure.

The DDE platform serves an important role as a bridge between DDE major scientific issues and general information technology infrastructures. On the one hand, it integrates DDE big knowledge and data, while providing efficient tools as well as practical systems for the construction of new data systems. On the other hand, it enables researchers to easily use cloud resources, build diverse scientific applications, and disseminate scientific research outcomes to decision makers, stakeholders, and the public.

4.3.2 The Key techniques of the DDE Platform

Cloud Native Platform High Concurrency Service Architecture

The DDE Platform adopts a cloud native high concurrency service architecture, ensuring the high availability of the entire link of platform information services and providing reliable support for high concurrency access and high-performance online computing.

The core of the architecture is consists of the three parts:

1) Cloud native base: through the idea of cloud container and microservice, all management modules of the service layer are deployed for multiple instances, and the system load resilience is improved through load balancing, elastic scaling, and other rules.
2) Data message government: message queue is used to process service requests asynchronously for message decoupling and peak clipping, achieving global high concurrency service.
3) Globalization Service: realize rapid response and global services through multi-level data cache, remote multi backup deployment, data compression, etc.

Unified Expression and Management of Multidisciplinary Analytical Model

Scientific computing engine of the DDE Platform realizes the unified encapsulation of various analysis models in different operating environments through the containerized model organization. The principle and operating environment of each tool are black-boxed. Therefore, the practical semantic expression of the model can be realized only by defining the primary operating conditions, such as the input and output specifications of the black box, the requirement of resources, and the intellectual property rights and usage agreement.

Construction, Collaboration, Publishing and Sharing of Model Chains

Online Creation: users can use the workflow creation page to create and manage a personal model chain workflow online.

Collaboration, Publishing, and Sharing: users can customize and modify the existing model or workflow of the system, choose whether to grant the permission and make the modified model or workflow public, promote the improvement of the model chain, forming a complete DDE workflow ecosystem.

Based on the unified specification of the containered analysis models, the DDE Platform has realized the construction, storage, collaborative update, release, and sharing of the workflow diagram of the model chain in line with the Common Workflow Language (CWL). It supports multiple people to cooperatively analyse the same complex big scientific research topics.

Model Chain Scheduling and Entire Process Monitoring

In terms of resource scheduling of the model chains, the DDE Platform organize the task units and resource units into groups at the logical level based on infrastructure virtualization, and realizes the physical execution optimization under different strategies based on different scheduling algorithms. At the same time, The logs of running container is monitoted and collected through deamon processes, processed with pre-defined rules and pushed to the downstream log analysis and display services. Therefore, users can get access to the whole lifetime of the overall analysis to get better understand of the research.

Online Multidimensional and Multiscale Visualization

In order to support the demand for visual and interactive analysis of calculation results in the exploratory analysis of scientific problems, the templated online real-time mapping of multi-source / multi-type complex geographic data in the network environment is realized, and pluggable data source plug-ins are provided to achieve the sharing and publishing of various geological data in various storage services on the cloud.

Linear Scalable Multi-scale Time Framework

In order to support data management and service in deeptime, the DDE Platform realizes a linear scalable multi-scale time frame. As a result, users can reconstruct and restore geological scenes, events, and elements based on various time axes, such as geological age time axis, chronostratigraphic time axis, geomagnetic polarity time axis, solar activity, and lunar geological time axis.

Four-dimensional Spatiotemporal Fusion Based Deep Time Earth Data Cube

From four-dimensional (x, y, z, t) spatiotemporal modeling and considering the needs of multi-level spatiotemporal scale measurement of digital earth, it is developed with six-dimensional (x, y, z, spatial scale as S, time pointer as TP, time scale as TS) spatiotemporal modeling, and the earth grid subdivision method is used to aggregate the three-dimensional expression of space (x, y, z) into one-dimensional lattice element (or volume element), constructing a space-time cube model of four-dimensional space-time fusion of deep time earth (cell or voxel as G, S, TP, TS).

4.3.3 The Core Engines Empowering DDE Platform

4.3.3.1 Geoscience Data Engine (DeepLake)

The DDE Platform needs to provide a unified data retrieval portal for various Earth science data sources, so that geoscientists can easily access large amounts of public Earth science data (including raw, combined, derived data and literature) online. The data subset, if permitted, can also be migrated to the storage system for analysis and visualization. The DDE Platform also needs to have the ability to manage private and proprietary data uploaded by users. Thus, we develop the Geoscience data engine.

Geoscience data engine provides data services such as data governance, data diversion, and data assets and realizes various functions such as multi-source and multi-modal space-time integration of various data, processing, retrieval, and matching.

Data Governance: the data governance module covers the whole life cycle of data, realizes the unified management of data through the monitoring and governance of the whole process from creation to extinction, and ensures the integrity, accuracy, consistency, and effectiveness of data in the whole process of collection, concentration, transformation, storage, and application. It mainly includes metadata, data standard, quality, data security, data life cycle management, and data map management.

Data Diversion: Data diversion aims to solve the problem of user sharing and acquisition and provide different data acquisition mechanisms according to different applications, including standard data access services, customized online retrieval and download, and 3D visualization support. Data streaming can distribute the data in the system to different applications in different forms, achieve the purpose of customization and rapid acquisition of data in the application layer, and quickly share the data, computing power, and intellectual resources of the system in the application layer. The system provides various data diversion modes, including multi-mode rapid retrieval of data, customized data download, map sharing, model sharing, and case achievement sharing. The data diversion module will build a unified data service bus, uniformly manage internal and external API services, support the access, query, and retrieval of business topics, portraits, and indicators, improve data experience and efficiency, and finally realize the realization of data assets. The data service adopts serverless architecture. It only needs to pay attention to the query logic of the API itself and does not need to care about infrastructure such as the operating environment. The data service will prepare computing resources and support elastic expansion. API Gateway provides API life cycle management, access control, flow control, and other functions. Flow control supports limiting the number of API calls from different dimensions such as user, application, and period to protect back-end services.z

Data Assets: with the advent of the era of big data, the importance of data has been raised to an unprecedented height, and "data is an asset" has been widely recognized. In the face of massive ample data assets, kinship analysis effectively ensures data fusion (aggregation). In addition, kinship can be used for data traceability, evaluation of data value and quality, and as a reference for data archiving and destruction.

4.3.3.2 Online Cooperate Scientific Computing Engine (DeepEngine)

DeepEngine is an online collaborative scientific analysis engine with unlimited cloud resources and a necessary export of the DDE ecosystem. DeepEngine makes researchers no longer need to care about infrastructure resource capacity but only pay attention to three core resources: data, model, and computing resources. Under the traction of scientific analysis problems, each resource unit is assembled and adjusted on demand by the core participants of ecological construction, researchers in all geological problem analysis links; finally, DeepEngine will incubate data-driven innovative research works.

As one of the specific footholds of the DDE ecosystem, the DeepEngine scientific analysis engine is a vital bearing tool for its incubation capacity for big scientific problems. It has the following functions:

1) It is a low threshold multi-user online platform.
2) It is an open environment for data/model/analysis process access and sharing so that the community's resources of the DDE ecosystem can be continuously and effectively supplemented.
3) It can support online collaboration of complex analysis processes, and different experts on demand can customize each link.

4) At the same time, it must also support online interactive exploration and analysis: users can directly obtain the data and analysis status of each link of the complex analysis process on the platform to realize the iterative development of the scientific problem analysis process.

Its technical framework is based on scalability, high-efficiency, reliability, and high-security cloud native service infrastructure. Two components of DeepEngine are Online Computing and Custom Workflow.

1) ***Online Computing:*** DeepEngine's online computing module is built on the scalability, efficiency, reliability, and high security of cloud native service infrastructure. These infrastructures include cloud container engines, microservices, DevOps, and service grids. Users can choose their own development environment through the online computing module to achieve online code operation. The user can use the intelligent computing module to create a new development environment, choose the container image (obtain the image from the mirror warehouse in the DeepEngine), select the environment type, define the environment name, environment variables, and complete the development environment. After the development environment is set up, users can upload the required data in the development environment, then write related codes and run online. The running results can be viewed in real time. Users are able to freely manage existing codes and can also use the classic codes preset in DeepEngine to build the development environment and run the codes online.
2) ***Custom Workflow:*** DeepEngine supports users to develop their own models and manage their own AI algorithms and data preprocessing tools. Users can quickly complete the layout of the established model through the process visualization layout page to quickly build an end-to-end solution. Specifically, the workflow module can be divided into three parts: model management, process management, and job management. *Model management* includes an intelligent calculation module to cover the function of model management, which is convenient for users to quickly manage their own independent AI algorithms and data processing tools. Users can use the Docker application container engine to complete the model code coding, make a model image and upload it to the platform, then complete the creation of the model on the platform and incorporate it into model management. In addition to supporting models developed by users themselves, DeepEngine provides preset models. Users can also directly use the preset models to complete the process orchestration without paying attention to model development. *Process management* provides easy-to-use and extensible process orchestration tools based on a drag-and-drop interface to achieve data modeling, analysis, and visualization, thus helping users to quickly build end-to-end solutions with zero coding. After the process is arranged, the user can start the process construction job and view the execution status in the *job management* module. The job management module provides job execution, full-process monitoring, and process visualization, and helps users deal with recalculating job tasks. Users can also perform job cloning operations, change job input parameters and environmental parameters, and realize rapid construction of multiple similar jobs.

4.3.3.3 Four-dimensional Earth Visualization Engine (DeepGlobe)

Based on the linear scalable multi-scale time frame and the global equal volume three-dimensional grid framework, the four-dimensional earth engine provides map services and 2D, 3D earth visualization services in RESTful style interface for deep time. Thus, the display and operation of massive multi-source heterogeneous geoscience data with an extended period are realized on the globe. The four-dimensional earth engine also provides map creation, grid customization, and other functions. It also provides intelligent spatiotemporal data processing environment and comprehensive data processing capacity for map data processors and cartographers to upgrade the quality of data products to meet the needs of terminal calls. In addition, it can provide web map product registration and release services and the DevOps guarantee.

Four-dimensional Earth Engine provides a data integration framework for surface and subsurface solid Earth data, spatio-temporal data processing, and a 3D geological body modeling tool for geoscientists to carry out sphere-coupling research. Four-dimensional Earth Engine also supports the four-dimensional digital reconstruction and spatiotemporal dynamic visualization engine for studies of the evolution of Earth materials, geography, life, and climate. At the same time, the Four-dimensional Earth Engine provides cartography, publishing and sharing services of full process and multi-directional support for the preparation, cartography, management, publishing and application of global 1:5 geological big data. Four-dimensional Earth Engine incorporates Three components:

1) ***DeepGlobe 4DBuilder:*** DeepGlobe 4DBuilder is the core module of DeepGlobe. Its objective is to construct a space-time framework Earth data. Based on various geological exploration data, geological analysis models, and geological expertise, the four-dimensional dynamic model of Earth from the Cambrian to the present (the time interval best constrained by data) will be constructed. Based on this model, each position of Earth in different geological ages can be inferred forward and backward.
2) ***DeepGlobe Visualization:*** The DeepGlobe Visualization engine is the core true 3D rendering engine that can support the dynamic visualization of time and space. The engine provides game-level realistic rendering capabilities, and supports high-quality real-time visualization of massive global geological boreholes, outcrop profiles, geological body models, and four-dimensional spatiotemporal scene data by calling the DDE spatiotemporal database.
3) ***DeepGlobe(DeepMap) Fusion:*** DeepMap Fusion series tools provide cartographers with intelligent spatio-temporal big data processing system environment and comprehensive data processing capability based on relevant standards, specifications and technical means, in terms of the quality of data products can be upgraded to satisfy the demand of terminal call. DeepMap Fusion includes data fusion processing tools, cartographic tools and product customization tools. DeepMap fusion is capable of connecting, fusion processing, mapping, network customization, operation and maintenance for DDE data. It mainly provides production, registration and publishing services for network map data products.

5. Geoscience Data Applied to Societal Needs

DDE can also be used to address many applied geological problems associated with energy, mineral, and water resources, thus contributing to societal needs and to the UN's Sustainable Development Goals (SDGs) which are a convenient set of objectives relating to societal needs.. To achieve these, DDE will need to collaborate with geological surveys, industry, and governments.

The SDGs are a collection of 17 global goals set by the United Nations General Assembly in 2015 for the year 2030 (Fig. 5-1).

Fig. 5-1. The 17 Sustainable Development Goals

All of the SDGs are linked in various ways and with varying degrees of complexity to geological science and digital geoscience.

An example is SDG7 'Ensure access to affordable, reliable, sustainable and modern energy' which aims at improving energy access, increasing renewables in the energy mix, energy efficiency, and integration and international cooperation.

Energy in its broadest sense enables business, industry, agriculture, transport, communications and modern services such as health care; but it also enables improvements in living standards. SDG7 is therefore intimately connected with most of the other SDGs (Stephenson, 2021) mainly through providing improved living standards, economic growth and activity, and improved environmental protection (Fig. 5-2).

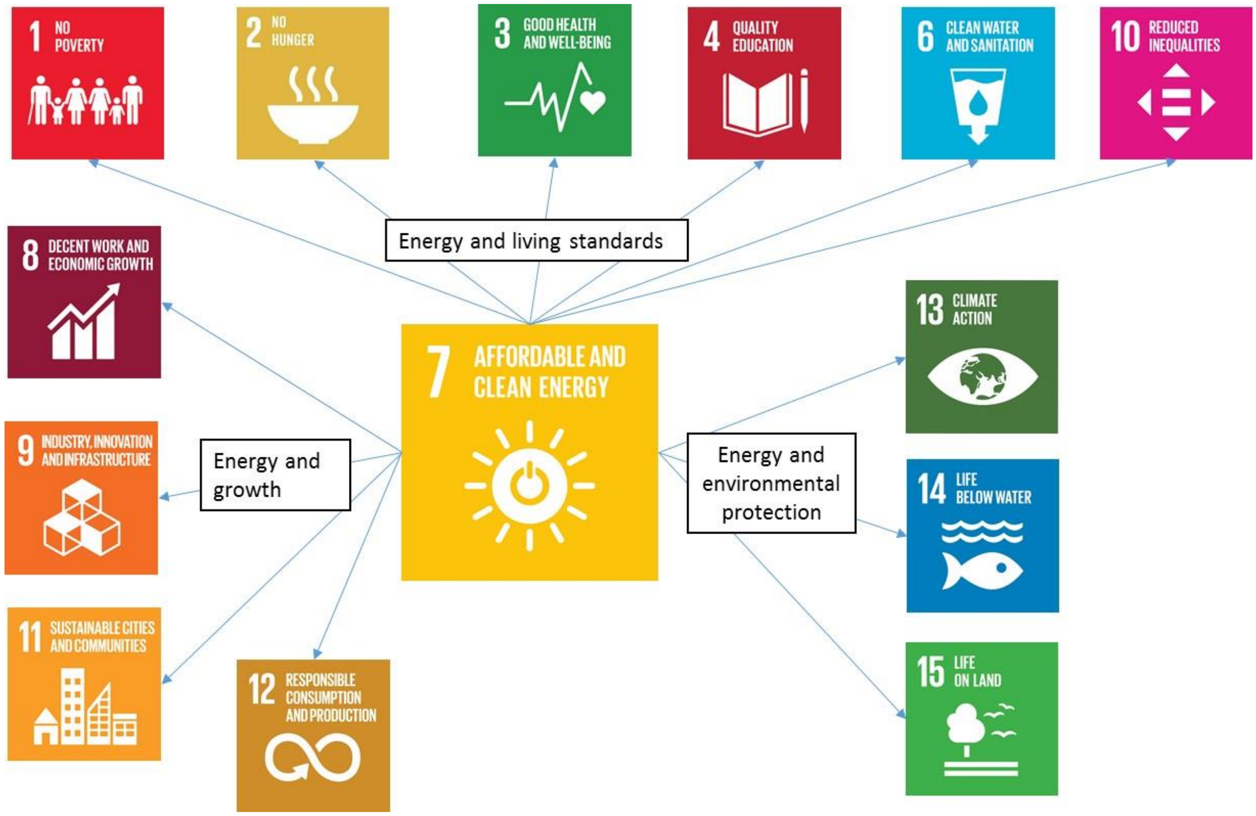

Fig.5-2. SDG7 and its connections with other SDGs

The services that energy provides improve human, social, economic and environmental conditions. Final energy use and the Human Development Index (HDI) are correlated (Steckel et al., 2013), the correlation implying early rapid gains in HDI with relatively small gains in energy usage, with HDI levelling off at levels of energy usage around 75 GJ/yr per capita.

On the negative side, energy (for example fossil fuel power and hydropower), can be produced and deployed in ways that pollute the environment, affect land use and increase greenhouse gas emissions. Similarly energy is an element of the food-energy-water nexus and thus its sustainability is tensioned against water and food (Stephenson, 2021). Finally, the financial value that energy can release, can also be siphoned into the ruling elites of kleptocracies and autocracies, rather than be cascaded down to benefit society at large. So the benefits of energy for sustainable development are strongly dependent on ethical governance and institutions (Torvik, 2009).

In general the effectiveness of energy systems to supply sustainable development depend on a number of factors (modified from Kaygusuz, 2012): (1) availability, affordability, security, reliability and safety of energy supplies; (2) environmental sustainability of the energy supply; (3) planning, design, construction, operation, financing and pricing of energy-using buildings, industrial processes, and transport systems in end-use sectors; (4) social and cultural norms of the use of energy; (5) access to alternative technologies and energy sources; (6) investment assistance to develop and deploy energy services; and (7) government policies that ensure that energy systems develop in a way that best supports and accords with sustainable development. These are summarised in Fig. 5-3.

Geoscience has a direct role in several of these areas including in establishing the geographical distribution, geological habitat, geotechnical feasibility of construction and infrastructure, and environmental sustainability, of energy supply.

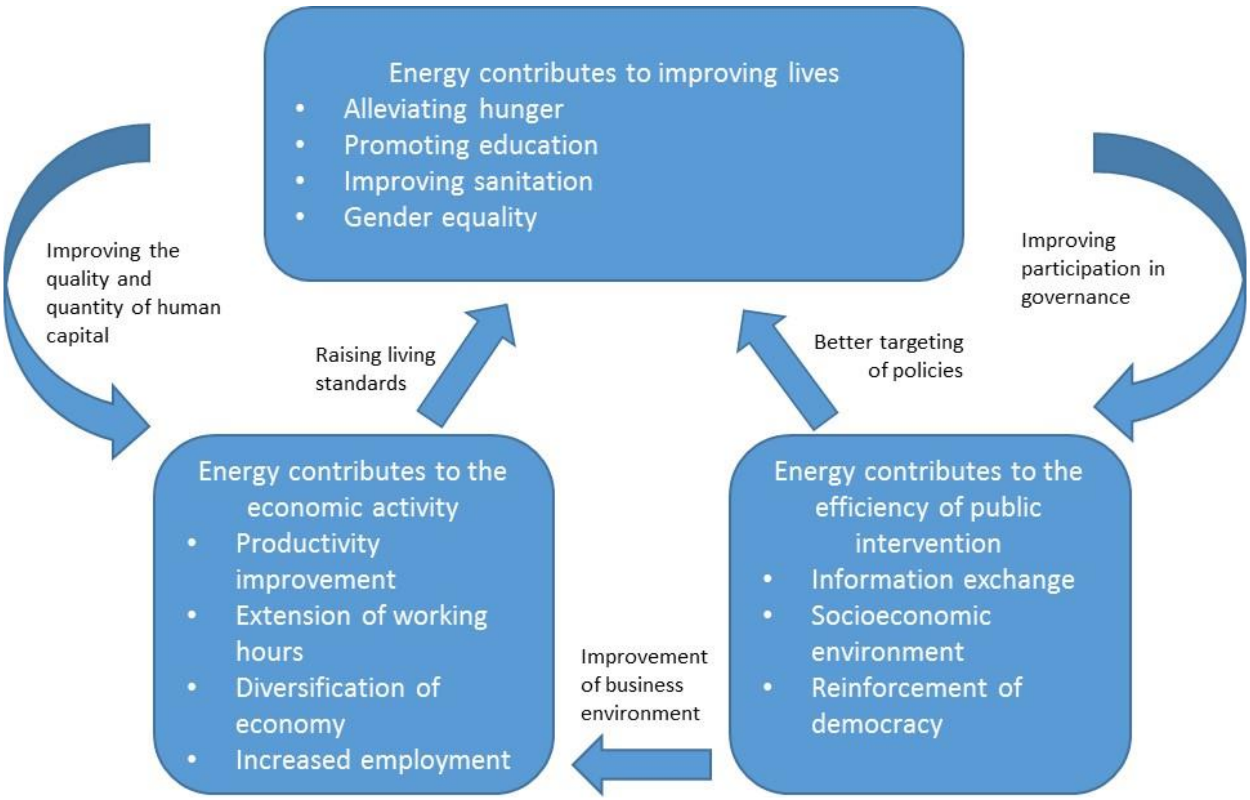

Fig.5-3. Energy and sustainable development(after Kaygusuz, 2012)

The 17 SDGs include 169 targets set by the UN, collectively aiming to end poverty; to ensure universal access to basic services (e.g., water, food, education, healthcare); to tackle diverse economic and social inequalities; to ensure sustainable consumption patterns; and to facilitate inclusive economic growth, social development, and environmental protection.

The SDGs require science research, innovation, capacity building, and technology transfer, including geological science, because understanding, monitoring, protecting, managing, and enhancing the natural environment are central to many of the SDGs. Knowledge of Earth's structure, processes, and resources, together with the ability to translate this knowledge into tools to inform policy and practice, can inform many key aspects of sustainable development. The activities required of geoscientists to deliver each SDG are different. Some will require the application of core geoscience knowledge, unique to the discipline. For example, ensuring access to water and sanitation for all (SDG 6) requires significant engagement by the geoscience community (e.g., hydrogeologists, geophysicists, hydrogeochemists) to understand and manage freshwater resources. Current knowledge, future research, and many of the practical skills required to meet this goal are significant strengths of the geoscience community. SDG 7'Ensure access to affordable, reliable, sustainable and modern energy,' has targets that are closely associated with geoscience, for example in exploration and feasibility studies for subsurface renewables such as geothermal, as well as sustainable use of fossil fuels within strict carbon budgets.

For other SDGs where subject matter is less geological, geoscientists will need to look at what actions they can take within their places of work. For example, geoscientists should take responsibility for championing and delivering gender equality (SDG 5) in those spheres that can be influenced.

Assignments of geoscience contributions to the 17 SDGs are shown in Table 5-1 (after Gill, 2017), with further analysis of the geodata implications of these assignments for the purposes of this White Paper.

Table 5-1 Assignments of geoscience contributions to the 17 SDGs by UNESCO

	Geology										Geoscience Deep-time data	
	Earth materials, processes and management								Skills and practice			
	Agrogeology	Climate change	Energy	Engineering geology	Geohazards	Geotourism	Hydro and	Minerals and rock	Education	Capacity building	Misc.	
GOAL 1: No Poverty												• Data correlating bedrock geology and soil fertility • Facies data for geological studies of basins and rock suitability for CCS, hydrogen storage, compressed air energy storage • Studies of deep-time paleoclimate to fed into modern GCMs • Facies data for geological studies of basins for predicting ground

	Geology										Geoscience Deep-time data	
	Earth materials, processes and management								Skills and practice			
	Agrogeology	Climate change	Energy	Engineering geology	Geohazards	Geotourism	Hydro and	Minerals and rock	Education	Capacity building	Misc.	
												conditions, building and infrastructure development and tunnelling • Deep-time geological data to feed int better understanding of modern hazards • Deep-time geological data to improve public appreciation and engagement in geological tourism • Facies data for geological studies of basins for predicting hydrogeological conditions, groundwater availability • Facies and tectonic data for models predicting mineral and rock material resource and distribution
GOAL 2: Zero Hunger												• Bedrock geology and soil fertility • Facies data for geological studies of basins for predicting hydrogeological conditions, groundwater availability • Facies and tectonic data for models predicting mineral and rock material resource and distribution
GOAL 3: Good Health and Well-being												• Bedrock geology and soil fertility • Facies data for geological studies of basins for predicting hydrogeological conditions, groundwater availability

	Geology										Geoscience Deep-time data	
	Earth materials, processes and management								Skills and practice			
	Agrogeology	Climate change	Energy	Engineering geology	Geohazards	Geotourism	Hydro and	Minerals and rock	Education	Capacity building	Misc.	
GOAL 4: Quality Education									■			• Deep-time data freely available to schools and universities
GOAL 5: Gender Equality											■	
GOAL 6: Clean Water and Sanitation							■					• Facies data for geological studies of basins for predicting hydrogeological conditions, groundwater availability
GOAL 7: Affordable and Clean Energy			■	■				■				• Facies data for geological studies of basins and rock suitability for CCS, hydrogen storage, compressed air energy storage • Facies data for geological studies of basins for predicting ground conditions, building and infrastructure development and tunnelling • Facies and tectonic data for models predicting mineral and rock material resource and distribution
GOAL 8: Decent Work and Economic Growth			■			■		■				• Facies data for geological studies of basins and rock suitability for CCS, hydrogen storage, compressed air energy storage • Facies data for geological studies of basins for predicting ground

	Geology										Geoscience Deep-time data	
	Earth materials, processes and management								Skills and practice			
	Agrogeology	Climate change	Energy	Engineering geology	Geohazards	Geotourism	Hydro and	Minerals and rock	Education	Capacity building	Misc.	
												conditions, building and infrastructure development and tunnelling • Facies and tectonic data for models predicting mineral and rock material resource and distribution
GOAL 9: Industry, Innovation and Infrastructure												• Facies data for geological studies of basins for predicting ground conditions, building and infrastructure development and tunnelling
GOAL 10: Reduced Inequality												
GOAL 11: Sustainable Cities and Communities												• Data correlating bedrock geology and soil fertility • Facies data for geological studies of basins and rock suitability for CCS, hydrogen storage, compressed air energy storage • Facies data for geological studies of basins for predicting ground conditions, building and infrastructure development and tunnelling

	Geology										Geoscience Deep-time data	
	Earth materials, processes and management								Skills and practice			
	Agrogeology	Climate change	Energy	Engineering geology	Geohazards	Geotourism	Hydro and	Minerals and rock	Education	Capacity building	Misc.	
		■		■	■	■	■	■				• Deep-time geological data to feed int better understanding of modern hazards • Deep-time geological data to improve public appreciation and engagement in geological tourism • Facies data for geological studies of basins for predicting hydrogeological conditions, groundwater availability • Facies and tectonic data for models predicting mineral and rock material resource and distribution
GOAL 12: Responsible Consumption and Production			■			■	■	■	■	■	■	• Deep-time geological data to feed int better understanding of modern hazards • Deep-time geological data to improve public appreciation and engagement in geological tourism • Facies and tectonic data for models predicting mineral and rock material resource and distribution
GOAL 13: Climate Action		■		■	■		■		■			• Studies of deep-time paleoclimate to fed into modern GCMs
GOAL 14: Life		■					■				■	• Studies of deep-time paleoclimate to feed into modern GCMs

	Geology										Geoscience Deep-time data	
	Earth materials, processes and management								Skills and practice			
	Agrogeology	Climate change	Energy	Engineering geology	Geohazards	Geotourism	Hydro and	Minerals and rock	Education	Capacity building	Misc.	
Below Water												
GOAL 15: Life on Land												• Data correlating bedrock geology and soil fertility • Studies of deep-time paleoclimate to fed into modern GCMs • Deep-time geological data to feed int better understanding of modern hazards • Facies data for geological studies of basins for predicting hydrogeological conditions, groundwater availability
GOAL 16: Peace and Justice Strong Institutions												
GOAL 17: Partnerships to achieve the Goal												

To illustrate in more detail how geoscience data will contribute to these aspects of the SDGs, a number of case studies are given.

Case study 1: Porphyry copper deposits

Contributes to SDGs 1,7, 8, 9, 10, 11, 12

Porphyry copper deposits (PCDs) are copper ore bodies that are formed from hydrothermal fluids that originate from magma chambers beneath the deposit. Porphyry copper deposits are currently the largest source of copper ore. Most of the known porphyry deposits are concentrated in western South and North America and Southeast Asia and Oceania - along the Pacific ‘Ring of Fire’. The highest concentration of the largest copper porphyry deposits is in northern Chile. Almost all mines exploiting large porphyry deposits produce from open pits. It is estimated that the Earth's porphyry copper deposits contain approximately 1.7×10^{11} tonnes of copper, equivalent to more than 8,000 years of global mine production. This copper will underpin industry in all of its forms and particularly the energy transition to electric vehicles because of its importance to electrical drive systems and batteries. The exploration and development of PCDs therefore contribute directly to societal needs.

PCDs are generated in continental arcs in response to plate convergence, subduction and collision. In general, PCD formation is associated with igneous activities from the upper mantle or lower crust during the evolution of continental lithosphere. Magmatism is often episodic with a varying magmatic addition rate and magmatic flare-ups due to deep rooted events. Key processes include partial melting in the mantle wedge, potentially causing density differentiation at the Moho. At shallow depth, magma interacts with country rock, hydrothermal fluids, and surface water systems, leading to supercritical fluids in a phase transition zone. Two hypotheses on the episodic behaviour of arc systems include external and internal control. External attributes include plate reconfigurations, changes in mantle flow, and in magma production etc. Internal attributes involve crustal thickening, melting, delamination, slab break off, roll back, slab split, and flattening (Fig. 5-4).

Cretaceous and Cenozoic PCDs are not random, but occur in distinct regional clusters. To investigate the clusters of PCDs and their governing factors requires analysis in various ways, including frequency distribution and size distribution, temporal distribution (i.e., number and sizes of deposits versus age), and spatial distribution along a profile parallel or perpendicular to the subduction zone.

In order to understand the governing factors of clustering, the PDC clusters are compared with abrupt changes or singularities of kinematical properties of plate motions, as well as geometrical properties of subducting slabs. Fig. 5-1 highlights the linking of temporal PCD clusters to change of plate velocity in the Andean subduction zone. Integration of various geodatabases through deep time in a paleogeographic references frame through DDE will allow us to link processes associated with subduction with mineralization processes in the crust. Data and interpretation have implications that will allow for a better understanding of the processes leading to formation (distribution, size) of PCDs, which are vital for modern industry, technology, and decarbonization. Linked databases and models can be used to gain insights into porphyry copper deposits that are not possible through simple analysis of single or pairs of databases. DDE will link georeferenced databases and models of this type so that they can be used more efficiently.

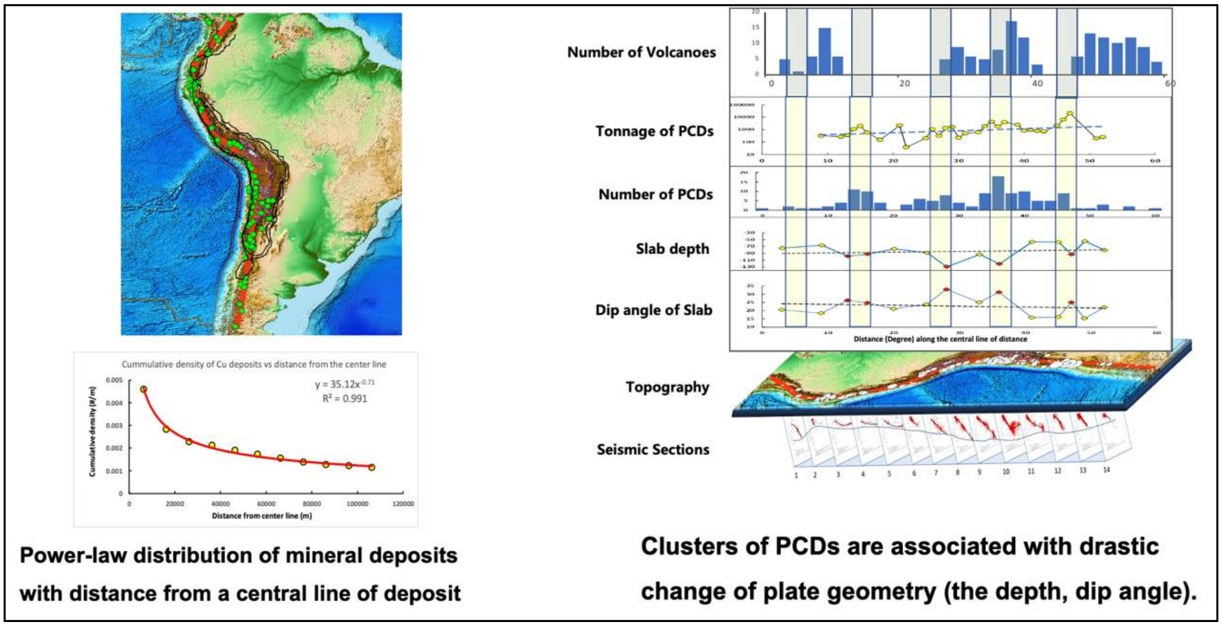

Fig. 5-4 Clusters of PCDs vs local change of subduction slab geometry in the Andes (Cheng, 2019)

Case study 2: African groundwater

Contributes to SDGs 1, 2, 3, 6, 7, 8, 9, 10, 11, 12, 13

Groundwater is the major source of drinking water in Africa. Demand will increase as Africa's economy and population grow. Despite its importance, there is still a dearth of detailed quantitative information on groundwater in the Continent. Groundwater storage is not often factored into assessments of freshwater availability, nor is an understanding of the aquifer as a reservoir, including its porosity, permeability, and chemistry, all of which will affect its water yield and storage capacity, and which are fundamentally governed by the vertical and lateral facies variation within the aquifer unit, which are in turn fundamentally governed by palaeogeographical and post depositional effects such as diagenesis. Understanding the geology of groundwater supply and storage will therefore contribute directly to societal needs, particularly as climate change begins to affect the amount and distribution of surface water.

Many developed countries, have sophisticated and detailed hydrogeological models of major aquifers (for example in south east England, the Cretaceous Chalk aquifer). But the facies variation and factors that control variation – –in African aquifers are not well known. Consequently, a number of questions are hard to answer in water-stressed regions that will likely become more water stressed due to climate change.

For example, predicted reductions in run-off in climate change forecasts have serious implications for populations in Africa. How much of that deficit can be compensated for by, for example, groundwater? Are the local aquifers large enough and will water wells make up the deficit without over-exploitation? Are the aquifers understood well enough as rock reservoirs for water? Will recharge in the area maintain the sustainability of the aquifer and the water wells or will it be affected by continuing climate change in the area? Similarly, what would the effects be of a large increase in run-off? What measures might have to be put in place to deal with flooding? Much of the research needed to answer these questions locally has yet to be done.

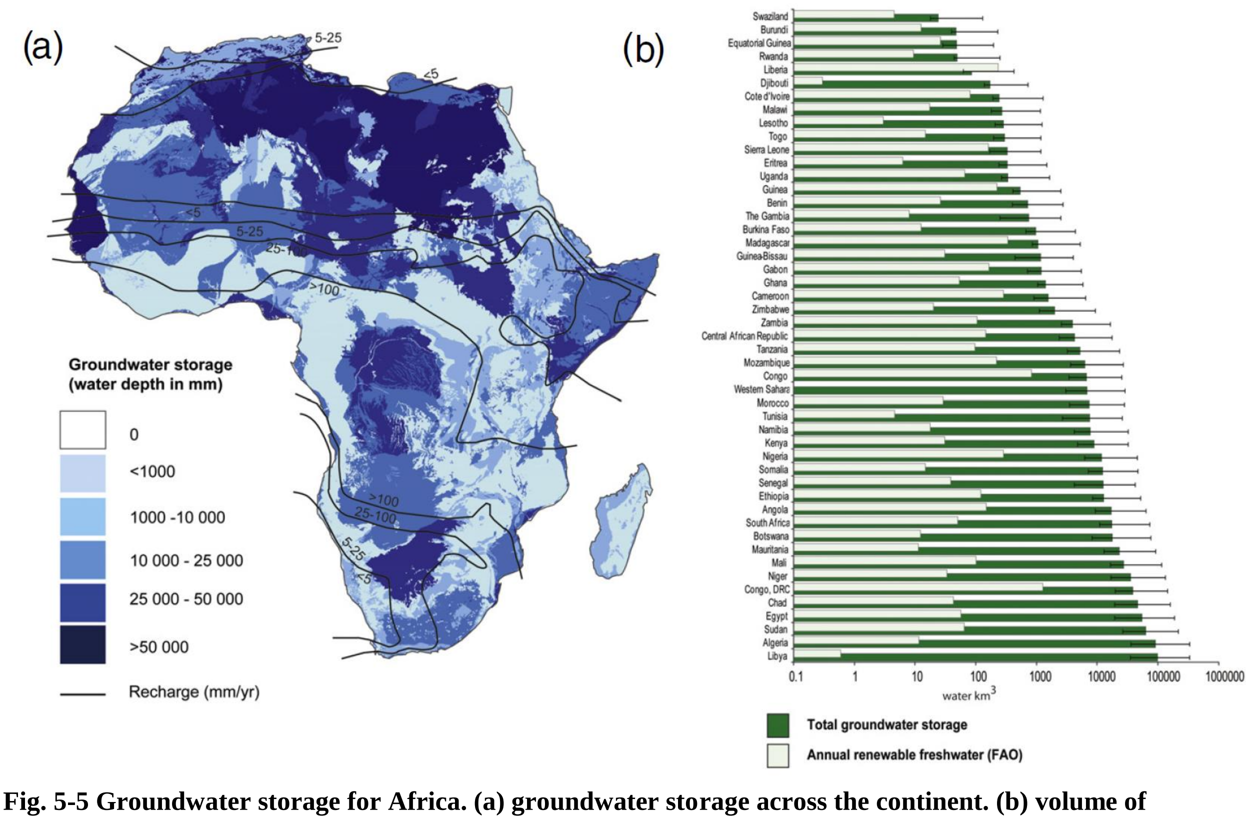

Fig. 5-5 Groundwater storage for Africa. (a) groundwater storage across the continent. (b) volume of groundwater storage for each country (MacDonald et al., 2012).

Linked databases to be developed by DDE, including hydrogeological models, recharge data, meteorological data, sediment flux, sub crop geology, basin subsidence, sequence stratigraphy, compaction, geomechanics, and tectonics data, will rely to some extent on a primary understanding of heterogeneity from palaeogeographical, basinal, and diagenetic studies, which will derive from deep-time datasets of the relevant stratigraphic units. These linked databases to be developed by DDE could lead to more accurate models for groundwater storage, including aquifer properties in African sedimentary basins. Such advances will underpin sustainable development in some of the poorest countries, and countries most vulnerable to climate change, in Africa (Fig. 5-5).

Case study 3: Permo-Triassic salt provinces of Europe: storage of heat and hydrogen

Contributes to SDGs 1,7, 8, 9, 10, 11, 12

Some major sedimentary basins host substantial natural resources. Many have been exploited already but others are poorly known or not yet discovered. Across Europe, Palaeozoic, Mesozoic, and Tertiary basins, in particular, provide access to groundwater, minerals, materials for infrastructure, and energy resources and storage. For example, the Permo-Triassic salt provinces of Europe, North Africa, and the Atlantic Margins provide potential large-scale underground storage opportunities for gas, such as hydrogen for a possible hydrogen economy, compressed air storage for energy storage to combat renewables intermittency (Fig. 5-6).

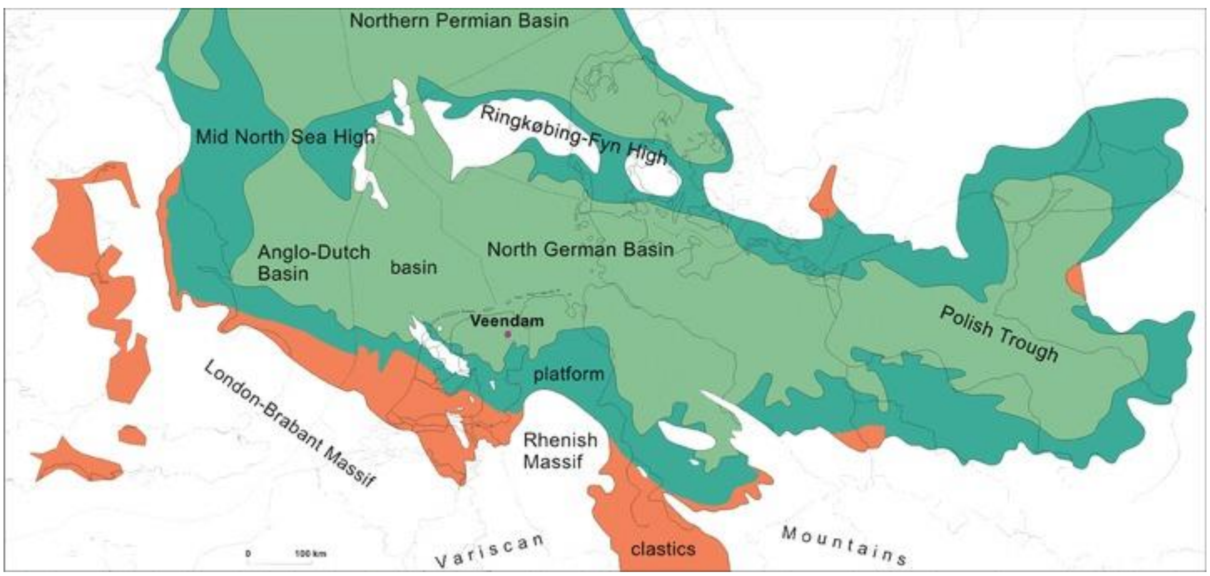

Fig. 5-6 Present day distribution and facies map of the Zechstein Group in the Southern Permian Basin. This will provide potential large-scale underground storage opportunities for compressed air and hydrogen (Drijkoningen et al., 2012).

Heterogeneities (facies, diagenesis, faults) govern the potential uses of lithological units within these basins. Therefore, integrated geo-referenced databases for sediment flux, sub-crop geology, basin subsidence, sequence stratigraphy, compaction, geomechanics, and tectonics can support simulation of basin development ('digital twin' basins) and 'forward stratigraphic modelling' to provide important information on the distribution of facies, diagenesis, and faulting in important resource basins. Linked databases in DDE will enable ‘digital basins’ to be developed to simulate the growth of basins and their heterogeneities and the management of basin usage (e.g. pressure management and groundwater management and control).

Case study 4: Bedrock geology and soil fertility: ‘hidden hunger’

Contributes to SDGs 1, 2. 3, 6 7, 8, 9, 10, 11, 12, 13

The UN Sustainable Developments Goals emphasise the need for improvements in health and wellbeing (SDG 3), and the removal of poverty (SDG 1) and hunger (SDG 2). Even where staple food is sufficient, ‘hidden hunger’ from mineral deficiencies (such as zinc, iron, iodine, calcium, and selenium) is prevalent. These also affect the wider SDGs by causing cognitive dysfunction, growth retardation, increased mortality, and disease (SDGs 4, 5, 10 and 15).

To understand ‘hidden hunger,’ a noted problem in parts of East Africa, we need to understand the way that the bedrock interacts with soil, and how the soil interacts with crops, such as the geochemical limits of soil–plant micronutrient transfer. However, understanding and integrating parent rock and soil geochemical processes at multiple scales for agricultural and health purposes has not yet been achieved in sub-Saharan Africa, due to data gaps and technical and analytical capacity constraints. Understanding soil geochemical processes is essential to support agricultural and public health policies (e.g. liming, organic residue incorporation, mineral deficiencies, and toxicities). Deep-time geological data and minerals data linked by DDE could contribute to soil analytical chemistry, geospatial integration and analyses, and soil management to understand ‘hidden hunger’ better.

6. Implementation Plan

The 10-year horizon of this Science Plan needs to be translated into a medium-term strategic framework for its implementation. The Governing Bodies of DDE may initiate, as early as possible, the drafting, review, and adoption of a Medium-Term Plan (MTP; over a 5-year period, for example) that defines an overarching goal, priority objectives, implementation processes, outcomes, and time-bound deliverables. In setting objectives, priority should be given to establishing DDE cyberinfrastructure critical to DDE's vision as a pioneer initiative dedicated to data-driven discovery. Other priorities could focus on: creation of a high-quality web-site that is DDE's voice and primary communication and outreach channel; developing a global network of Research Centers of Excellence (RCEs); designing new, harmonized databases and/or innovative combinations of existing databases to address critical questions on the 4E themes (evolution of materials, geography, life, and climate) and geosciences for societal benefits; and steps and targets for mobilizing funds for the full implementation of the MTP and this Science Plan during the years after the MTP until 2029.

The following sections address certain essential features of the implementation process related to:

1) DDE's relationship to other Geodata initiatives;
2) DDE cyberinfrastructure construction;
3) Program governance and management;
4) Funding.

6.1 Relationship of DDE to other Geodata initiatives

DDE covers all the world regions, the whole spectrum of geoscience disciplines and themes such as energy, minerals, groundwater, and evolution of land and seascapes and their biodiversity where geoscience linkages to societal benefits could be demonstrated. Uniquely, it will harmonize deep-time data, mainly in the 'long tail' Earth record with contemporary digital data; in this sense, DDE is different from other digital initiatives that primarily focus on collecting and managing data for designing current Earth system observation and monitoring initiatives that primarily are aimed at prediction of hazards (Fig. 2-1). The added value of DDE will derive from harmonizing existing databases covering deep-time that are, at present, in disparate distributed sources, many of which are built mainly on data from contemporary sensors and observations.

The new digital infrastructure to be developed under DDE will make discovering, integrating, sharing, managing, and utilizing deep-time Earth data possible and highly interesting and challenging. DDE will establish a global deep-time digital Earth data sharing alliance (hereafter referred to as DDE Data Alliance, or DDE/DA) to enable orderly sharing of such data around the world. Currently, there are as many as 84 geosciences databases (Appendix 1). As shown in the last column of Appendix 1, DDE will aim to establish links with the top 12 of the 84 as the immediate priority in developing DDE/DA.

The purpose of DDE/DA is to establish a global data sharing network, so that deep-time Earth data can be discovered, accessed, utilized, and recycled (reused) to help provide solid, open-data support for a global deep- time discovery paradigm that will significantly transform Earth sciences research. DDE/DA will set itself a goal to discover and integrate all existing deep-time Earth data, and generate more by deploying new information technologies, and ensure their global accessibility through an open-science, data-sharing network. As evident from Appendix 1, all of the 84 databases are operated from developed economies and China; DDE/DA, in building the data sharing network will make data resources of these global and regional network accessible to all countries, in particular less developed nations.

DDE/DA will be open to all Founding members of DDE, international organizations, research institutes, educational entities, database operators, scientists' teams, and data volunteers. Members can publish or integrate global deep-time Earth data through the DDE big data platform. The rights and obligations of DDE/DA participants will be defined as part of work to be done for the construction of the DDE cyberinfrastructure.

6.2 Program Governance, Management and operations

Governance defines the structure and processes that exists in and between formal institutions that allows them to work together for a common goal. Good governance means that all stakeholders buy into the initiative, recognize its benefits, and feel part of its decision-making processes. Good governance is critical to a well-functioning DDE Program.

DDE will assemble a wide range of stakeholders and have strong internal governance mechanisms to ensure efficient decision making and effective management and operations. The governance structure will build trust and foster active engagement and participation of members, partners, and other intended beneficiaries; it will promote gender and geographical diversity in its decision-making, advisory, management, and operational organs (Fig. 10)

6.2.1 Management and Operations of the DDE Program

The management of DDE will be guided by the DDE Statutes and Bylaws. The governing bodies of the DDE Program form a hierarchy (Fig. 5-4), comprising the DDE Governing Council (GC) and Executive Committee (EC), Scientific Committee (SC), Secretariat and Working Groups (WG), and Task Groups (TG) [often referred to as WTGs in combination]. The DDE Advisory Board (AB) guides the work of the GC but does not have decision making or executive authority. Together, they will ensure effective governance, management, and operations of the DDE. The members of all organs described in Fig. 5-4 are elected for a four-year term. The following is a brief description of each organ.

6.2.1.1 DDE Board (DDE-B)

The DDE Board gives advice on, and supports the strategic aims of, DDE. It can set key performance indicators (KPIs) and suggest procedures for DDE performance evaluation. DDE-B Members are representatives of international (IUGS) and intergovernmental (UNESCO) organizations and reputed scientists and scholars who provide DDE with credibility in the international arena.

6.2.1.2 Governing Council (GC)

The Governing Council (GC) provides governance, offers financial and scientific overviews, makes major policy decisions, and reviews applications for DDE membership. Founding and other members who have formally joined DDE are eligible to nominate a representative to the GC. The GC Chair will be elected every four years. The Chair has the authority to appoint, from among the GC Members, a maximum of two Vice-Presidents during his/her 4-year tenure. The current Chair of the GC is Prof. Mike Stephenson, Executive Chief Scientist for Decarbonization and Resource Management at the British Geological Survey. Prof. Qiuming Cheng, IUGS President (2017-2020) will serve as an Ex-Officio Member of the GC.

6.2.1.3 Executive Committee (EC)

The Executive Committee, comprising the Chair, DDE Secretary General, Treasurer, and two Counsellors, manages the DDE Program and reports annually to the GC on operational activities, providing a budget report and a summary of scientific outcomes of DDE projects, activities, and initiatives. EC will exercise direct oversight of Secretariat affairs and day-to-day management of the DDE Program. Members of the EC will be elected every four years. The current Chair of the EC is Academician Chengshan Wang, Professor of Geology and Sedimentology, China University of Geosciences, Vice president of IAS.

6.2.1.4 Scientific Committee (SC)

The SC will include members by invitation only. SC members will all be internationally renowned experts from the research fields of interest to DDE. The Chair and members of the SC will be nominated by Founding Members and other organizations represented in the GC and AB. SC will carry out thorough scientific evaluations of project proposals and reports and it will establish criteria and conditions for allocation of DDE funds for projects on a competitive basis. The SC membership will be renewed regularly with each individual member serving SC for a period of four years. The current Chair of SC is Roland Oberhänsli, Professor Emeritus of Petrology and Geochemistry, Potsdam University (Germany), and Past President of IUGS.

6.2.1.5 Working Groups (WG)

DDE will include a number of Working Groups, which will provide the Executive Committee with scientific support, operational capabilities, and infrastructure to manage the program and to support DDE projects as determined by individual terms of reference of the Working Groups. Working groups may include both geo and data sciences themes. WGs provide the scientific and disciplinary foundation of the DDE Program and will play a significant role in the implementation of the Science Plan described in this document. A WG is proposed by the EC and approved by GC.

6.2.1.6 Task Groups (TG)

Task Groups will explore databases through projects and/or workshops. Work of some individual Task Groups, for example that on "Standards", will be of significance for DDE's emphasis on establishing FAIR databases. A TG is proposed by the EC and approved by GC.

6.2.1.7 Administrative support

The Kunshan City Government in Suzhou, China will host the DDE International Secretariat. The Secretariat will comprise both Chinese and International staff. The Secretariat will serve DDE for a 5-year term; the location and host of the Secretariat will then be chosen through an open invitation process.

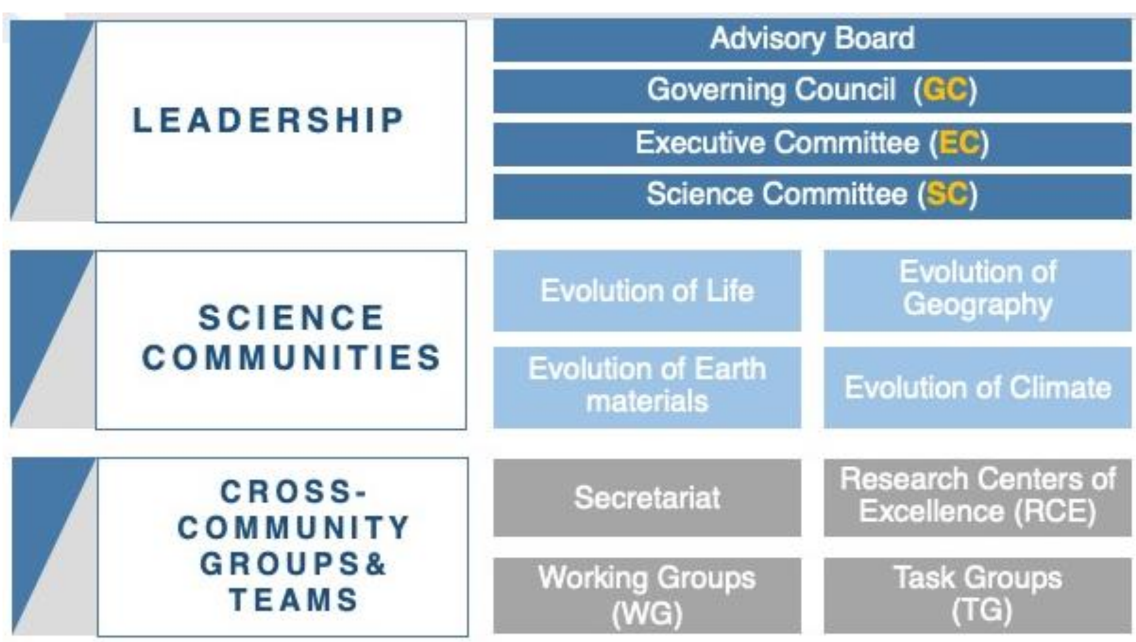

Figure 5-4: Structure of DDE

6.2.1.8 Members and Partners

DDE operations, including data-driven research, organization of scientific seminars and symposia on topics of global interest, and producing reports and publications, will be done in collaboration with a range of Members and Partners.

Members formally join DDE and sign a Memorandum of Agreement (MOU) to abide by DDE Statutes and Bylaws. Partners are organizations that do not seek formal DDE Membership but collaborate with DDE is areas of mutual interest. All organizations joining DDE before the end of 2021 will be recognized as DDE Founding Members. At time of writing of this document, there are 21 Founding Members:

American Association of Petroleum Geologists (AAPG)
British Geological Survey (BGS)
China Geological Survey (CGS)
Coordinating Committee for Geoscience Programmes in East and Southeast Asia (CCOP)
Commission for Geoscience Information (CGI)
Commission for Geological Map of the World (CGMW)
Geological Survey of India (GSI)
Korea Institute of Geoscience and Mineral Resources (KIGAM)
International Commission on Stratigraphy (ICS)
International Association for Mathematical Geosciences (IAMG)
International Association of Geomorphologists (IAG)
International Association of Sedimentologists (IAS)
International Association on the Genesis of Ore Deposits (IAGOD)
International Lithosphere Program (ILP)
International Palaeontological Association (IPA)
Russian Federal Geological Foundation (FBGU)
Russian Geological Research Institute (VSEGEI)
International Association for Engineering Geology and Environment (IAEG)
Network of Young Earth Scientists (YES)
International Mineralogical Association (IMA)
International Association of Hydrogeologists (IAH)

Interested organizations could apply and become DDE Members anytime. The process for applying for Membership, rights and obligations of Members, and other related information will be detailed in the DDE Statutes.

As an IUGS Big Science initiative, DDE will maintain close relationships with several GeoUnions and associations of IUGS. It will also develop close relations with other non-governmental (for example, IUGG) and inter-governmental scientific institutions, such as UNESCO. DDE provides an opportunity for open partnership between scientific unions, their bodies and affiliated members, and regional and international organizations, which include, but is not limited to, scientific associations, universities, geological surveys, publishers, data centers, and others who share DDE's vision to create a data-driven research paradigm in Earth sciences.

Given the overlap of the time frame of this Science Plan (2020-2030) and the UN Agenda 2030 to deliver Sustainable Development Goals (SDGs), DDE has the potential to attract several UN organizations and their partners to launch collaborative initiatives. The timeframe of this White Paper also overlaps with that of several UN Decades; e.g., on Ocean Sciences for Sustainable Development (2020-2030), Ecosystem Restoration (2020-2030), Water for Sustainable Development (2018-2028), and Sustainable Energy for All (2014-2024). DDE's contributions to UN SDGs relating to water, oceans, energy, and land could therefore attract some organizations collaborating under the framework of these UN Decades, which may eventually seek membership or other collaborative relationships with DDE.

6.3 DDE Funding

DDE budgetary and financial management procedures must be transparent and conducted in accordance with international accounting standards in order for DDE to be able generate the necessary financing for implementation of this Science Plan. The MTP for the first 5 years of this Plan must define procedures and targets for raising the funds necessary for the implementation of the MTP, as well as for continuing implementation of DDE until 2029.

The Kunshan City Government, Suzhou, has committed support for the DDE Global Secretariat and the establishment of the first DDE Research Centre of Excellence (DDE/RCE) in Suzhou, China. Funding for the recruitment of an international Director and a certain number of post-doctoral fellows for DDE/RCE in Suzhou has also been secured. DDE/RCE in Suzhou has also mobilized financial support US$ 5,000/year for each DDG working group for the year 2020 and the financing of 4 enabling projects, each lasting over 2-3 year-period beginning from 2021, at approximately US$ 50,000 per year. The China Geological Survey (CGS), one of the Founding Members of DDE, is financing a project to establish a DDE Country Node for China with potential for cooperation for capacity building in other geographical regions.

Opportunities for continued fund-raising and project-financing in China are likely to remain encouraging. The challenge for DDE is to increase its visibility, as well as delivery of high-value products and services that will attract international donors, both from the public and private sectors, as well global funds, foundations, individuals, and organizations with an interest to support data-driven science. The benefit of the project to the potential funding entities must be clear, and it is important to avoid ambiguities with respect to intellectual property.

Funding and financial support for the implementation of this Science Plan may assume different modalities:

1) direct "unearmarked" funds provided to the DDE Secretariat for use to support Secretariat, administrative, and/or scientific studies to be determined by DDE EC/GC;
2) "earmarked" funds for defined uses, for example data-driven projects that address one or more UN SDGs or one of the 4E themes;
3) direct sponsorship of human, technological, data, and financial needs of one or more projects by an external partner who agrees to report to, and respond to feedback from, DDE EC/GC;
4) revenues generated via Membership fees, as well as sale of products and services generated and owned by DDE through its projects and initiatives.

It would also be worth exploring the possibility of working with a partner or set of partners who are global organizations. In this way, funding for projects could be held in an international fund and audited annually. A DDE Fund, or account, under a global organization could attract international donor and partner financing, including from private sector organizations in energy, minerals, natural resources, and data management sectors. A DDE Trust Fund set up under the auspices of a UN Body or in the country of origin of one of the Founding Members, with transparent financial management and communication frameworks, could be a way to build up the long-term financial foundation of DDE, particularly for project-specific financial contributions. The setting up of a such a Trust Fund, however, must precede feasibility and other studies that can inform DDE EC/GC on the probability of success for the establishment and sustainability of the Fund.

6.4 Outreach, education, and public affairs

Besides developing technologies and innovations, DDE also plans to engage in outreach, education and public affairs, which is also crucial for a scientific program.

Outreach: The partners of DDE will publish results regularly in international peer reviewed journals and present them at international science meetings, such as the IGC, AGU and EGU, as well as special meetings of professional associations in special symposia and workshops. The work and achievement of DDE will also be disseminated through other news outlets of Founding Members and Partners as well as popular science publications, e.g., National Geographic, New Scientist etc., to inform the broader scholarly and scientific community. New international programs such as DDE need to increase public awareness and education, international collaboration, transdisciplinary research and open data sharing in the Big Data world. Therefore DDE will aim to collaborate with popular geoscience centers, for example, with museums to build a digital version of the geological history of dinosaurs; it will interact with leading public educational initiatives via the web and other media to highlight lessons that we can extract from the earth's evolution of climate, geography, life and materials to guide our plans and strategies for planetary futures. It will produce a new digital Earth visualization by integrating all Earth layers with high-resolution deep-time, and four-dimensional space-time coupling. These efforts will be coordinated with partners such as the IUGS Commission on Geoscience Education, Training & Technology Transfer (COGE) and shared throughIGGP, which includes both the International Geosciences Program, a continuation of the UNESCO-IUGS International Geological Correlation Program (IGCP) that will commemorate it's 50-year anniversary in 2022, and the Program on UNESCO Global Geoparks.

In its participation in global, regional and national conferences DDE will choose the most appropriate cost-effective method of disseminating the outcomes of DDE projects and initiatives. They may include organizing special events, establishing booths and other information and data sharing outlets at conference venues, convening on-line webinars open to conference participants etc. Scientists belonging to working and task groups of DDE will attend events that directly relate to their respective disciplines and research themes and disseminate information on DDE projects and initiatives and their results and outcomes. DDE will also collaborate with its Members and Partners to make the best use of their combined on-line and print outreach channels to share the outcome of the work of DDE. DDE will also place particular importance to be present and share information, data, knowledge and experience at events that cut across several earth sciences disciplines and that could contribute to addressing current-day global environmental and developmental challenges. An example, is the annual URTeC (Unconventional Resources Technology Conference); one of the three sponsors of this event, the American Association of Petroleum Geologists (AAPG), is a Founding Member of DDE. DDE will also aim to organize special webinars or conferences and symposiums to promote its vision for strengthening the data driven research paradigm for earth sciences. It will aim to bring together data from multiple disciplines to enable UN Member States reach sustainable development goals, particularly by focusing on themes important for the global transition towards low carbon futures; examples include critical minerals for the green-energy agenda; carbon capture, use and storage (CCUS) in climate change mitigation and adaptation etc.

Education: DDE's most significant contribution to education will be targeted to tertiary educational institutions, including spreading knowledge among graduate students and post-doctoral researchers of the challenges and rewards of pursuing data-driven research and teaching/learning approaches. The goal of DDE educational activities is to raise awareness and competency to enable effective usership of data platforms, products, and research results. DDE's cyberinfrastructure team will comprise the following components that will provide specific products and services derived from the applications of tools and techniques of data sciences, i.e., AI, machine learning, big-data analytics, data mining, natural language processing, semantic web etc., to strengthen research and educational skills and competences of higher-educational institutions.

DDE will also train public of how to use DDE system. (1) Educational efforts will instruct individuals how to use the DDE Platform and how to make best use of its products and services to meet their own research and educational needs. In addition, guidance on how to use the platform for both individual and collaborative (group) projects will be provided. (2) DDE will put effort on education of how to contribute, use, and evaluate the various repositories, tools, and project-based uses of the data. Priorities will be given to projects that use real-world data for results that have useful implications for planning and policy-making. This group will play an important role in generating knowledge-graphs, a visually synthetic representation of the complexity of data accumulated through geological time, for several earth sciences disciplines and explore their use as educational, awareness raising tools. (3) Education of DDE data will allow users be formiliar with the data and offer them best way to discover, access, utilize and recycle data for research, educational and awareness raising purposes.

Public Affairs: DDE's focus on data-driven research could be of significant interest to public policy professionals, journalists and other communities who confront the challenges of linking data and science to decision making for the public good. The role of data and science are critical in building models and visualizing scenarios for the future whether to address public health measures to mitigate the impacts of pandemics or to prepare nations and communities to adapt to a warming planet and receding coast-lines..

Globally, climate change scenarios, for example, modelling the impacts of global warming with ice-cap melting and a related rise in the level of the oceans, are data-driven. With clear modeling and simulation of the impact on coastal areas with sea level changes, governments, municipalities, and advocacy groups can design their specific responses and strategies. Similar, data-driven modelling and scenario visualization could be performed by DDE working and task groups, members and partners for specific issues and questions that would be of public interest within the context a number of UN and other international/global agendas.

The most subscribed UN Agenda among nation states is the UN 2030 agenda to deliver on Sustainable Development Goals (SDGs). At the time of writing this document, the world is nearing the last three years of the UN Decade on Sustainable Energy for All (2014-2024) and entering year 4 of the UN Decade for Action on Water for Sustainable Development (2018-2028); 2021 will see the launch of two other UN Decades which last till 2031; i.e., one on Ocean Sciences for Sustainable Development and the other on Ecosystem Restoration. Some issues that could be of interest to the UN Decade for Sustainable Energy for All, such as critical minerals for green energy futures and CCUS, have been referred to earlier in this section. Other applied science questions raised in previous sections, for example, groundwater resources in Africa will be of interest to UN Decade on Action on Water for Sustainable Development.

DDE will continue to record videos to educate the public about its mission and vision, and potentially to offer workshops on how to use the open access parts of DDE. DDE has the opportunity to educate public professionals, journalists and other groups that serve as communicators of the role of science in public affairs about data, models, scenarios and other tools and techniques to imagine futures based on the knowledge of the past. In this effort, DDE must aim to illustrate new insights and knowledge that a data-driven approach to solid-earth sciences could contribute to addressing contemporary problems and issues linked to global warming, sea-level rise, sustainable development etc. DDE may organize special events on data-driven research paradigm at UN and other global forums targeting public policy and affairs professionals; the DDE/RCEs, the first in China nearing full development and others in US, UK and Russia under consideration, could be venues for introducing public policy and affairs personnel to data-driven research and their role in policy and decision making for the global commons and public good.

References:

Anbar, A.D., 2008. Elements and evolution. Science 322, 1481–1483.

Ansari, G.A., Saha, A., Kumar, V., Bhambhani, M., Sankaranarayanan, K., Chakrabarti, S., 2019. Neural Program Induction for KBQA Without Gold Programs or Query Annotations. Presented at the IJCAI, pp. 4890–4896.

Ansari, S., Del Greco, S., Kearns, E., Brown, O., Wilkins, S., Ramamurthy, M., Weber, J., May, R., Sundwall, J., Layton, J., 2018. Unlocking the potential of NEXRAD data through NOAA's Big Data Partnership. Bulletin of the American Meteorological Society 99, 189–204.

Boucot, A.J., Xu, C., Scotese, C.R., 2013. Phanerozoic paleoclimate: An atlas of lithologic indicators of climate, Phanerozoic Paleoclimate. SEPM (SOCIETY FOR SEDIMENTARY GEOLOGY), Tulsa, Oklahoma.

Bowles, J., Howie, R.A., Vaughan, D., Zussman, J., 2011. Rock-Forming Minerals, Volume 5A: Non-Silicates: Oxides, Hydroxides and Sulphides, Geological Society of London. Geological Society.

Cheng, Q., 2019. Integration of Deep-time Digital Data for Mapping Clusters of Porphyry Copper Mineral Deposits. Acta Geologica Sinica-English Edition 93, 8–10.

Diepenbroek, M., Grobe, H., Reinke, M., Schindler, U., Schlitzer, R., Sieger, R., Wefer, G., 2002. PANGAEA—an information system for environmental sciences. Computers & Geosciences 28, 1201–1210. https://doi.org/10.1016/S0098-3004(02)00039-0

Drijkoningen, G., van Noort, G., Arts, R., Bullen, J., Thorbecke, J., 2012. Seismic time-lapse effects of solution salt mining–a feasibility study. Geophysical prospecting 60, 239–254.

Fan, J., Hou, X., Chen, Q., Melchin, M.J., Goldman, D., Zhang, L., Chen, Z., 2014. Geobiodiversity Database (GBDB) in stratigraphic, palaeontological and palaeogeographic research: graptolites as an example. GFF 136, 70–74. https://doi.org/10.1080/11035897.2014.880070

Fan, J., Shen, S., Erwin, D.H., Sadler, P.M., MacLeod, N., Cheng, Q., Hou, X., Yang, J., Wang, X., Wang, Y., Zhang, H., Chen, X., Li, G., Zhang, Y., Shi, Y., Yuan, D., Chen, Q., Zhang, L., Li, C., Zhao, Y., 2020. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. Science 367, 272–277. https://doi.org/10.1126/science.aax4953

Ferebee, M.T., Cordner, D.E., Ritchey, N.A., Hunt, L.A., Piatko, P., Haberer, S.J., Wang, F.Y., 2007. Finding and accessing data at the NASA atmospheric science data center. Presented at the 2007 IEEE International Geoscience and Remote Sensing Symposium, IEEE, pp. 4009–4011.

Gill, J.C., 2017. Geology and the sustainable development goals. Episodes 40, 70–76.

Gray, J., 2009. Jim Gray on eScience: A transformed scientific method, in: Hey, T., Tansley, S., Tolle, K. (Eds.), The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Redmond, WA, pp. xvii–xxxi.

Gregory, D.D., Cracknell, M.J., Large, R.R., McGoldrick, P., Kuhn, S., Maslennikov, V.V., Baker, M.J., Fox, N., Belousov, I., Figueroa, M.C., Steadman, J.A., Fabris, A.J., Lyons, T.W., 2019. Distinguishing ore deposit type and barren sedimentary pyrite using laser ablation-inductively coupled plasma-mass spectrometry trace element data and statistical analysis of large data sets. Economic Geology 114, 771–786. https://doi.org/10.5382/econgeo.4654

Hannisdal, B., Peters, S.E., 2011. Phanerozoic earth system evolution and marine biodiversity. Science 334, 1121–1124. https://doi.org/10.1126/science.1210695

Haq, B.U., Al-Qahtani, A.M., 2005. Phanerozoic cycles of sea-level change on the Arabian Platform. GeoArabia 10, 127–160. https://doi.org/10.2113/geoarabia1002127

Hazen, R.M., 2014. Data-driven abductive discovery in mineralogy. American Mineralogist 99, 2165–2170.

Hazen, R.M., Downs, R.T., Eleish, A., Fox, P., Gagné, O.C., Golden, J.J., Grew, E.S., Hummer, D.R., Hystad, G., Krivovichev, S.V., 2019. Data-driven discovery in mineralogy: Recent advances in data resources, analysis, and visualization. Engineering 5, 397–405.

Heidorn, P.B., 2008. Shedding light on the dark data in the long tail of science. Library trends 57, 280–299.

IPCC, 2021. Summary for Policymakers, in: Masson-Delmotte, V., Zhai, P., Pirani, A., Connors, S.L., Péan, C., Berger, S., Caud, N., Chen, Y., Goldfarb, L., Gomis, M.I., Huang, M., Leitzell, K., Lonnoy, E., Matthews, J.B.R., Maycock, T.K., Waterfield, T., Yelekçi, O., Yu, R., Zhou, B. (Eds.), Climate Change 2021: The Physical Science Basis. Cambridge University Press.

IUGS, 2019. IUGS Annual Reports 2019.

IUGS, 2018. IUGS Annual Reports 2018.

Jackson, I., 2008. OneGeology: from concept to reality. Episodes 31, 344–345.

Jiang, J., Li, C., Hu, X., 2020. Advances on Sedimentary Database Building and Related Research: Macrostrat As an Example. Geological Journal of China Universities 26, 27–43. https://doi.org/10.16108/j.issn1006-7493.2019102

Kaygusuz, K., 2012. Energy for sustainable development: A case of developing countries. Renewable and Sustainable Energy Reviews 16, 1116–1126. https://doi.org/10.1016/j.rser.2011.11.013

Lehnert, K., Markey, K., Johansson, A., Ramdeen, S., Song, L., Devendran, S., 2019. Best practices for curating geochemical and petrological data: the EarthChem library. Presented at the AGU Fall Meeting Abstracts, pp. IN23D-0904.

Lehnert, K., Su, Y., Langmuir, C.H., Sarbas, B., Nohl, U., 2000. A global geochemical database structure for rocks. Geochemistry, Geophysics, Geosystems 1.

Li, X., Cheng, G., Wang, L., Wang, J., Ran, Y., Che, T., Li, G., He, H., Zhang, Q., Jiang, X., Zou, Z., Zhao, G., 2021. Boosting geoscience data sharing in China. Nature Geoscience 14, 541–542. https://doi.org/10.1038/s41561-021-00808-y

Lin, D., Crabtree, J., Dillo, I., Downs, R.R., Edmunds, R., Giaretta, D., De Giusti, M., L'Hours, H., Hugo, W., Jenkyns, R., 2020. The TRUST Principles for digital repositories. Scientific Data 7, 1–5.

Loveland, T.R., Dwyer, J.L., 2012. Landsat: Building a strong future. Remote Sensing of Environment 122, 22–29.

MacDonald, A.M., Taylor, R.G., Bonsor, H.C., 2012. Is there sufficient water to support the intensification of agriculture from 'land grabs'?'. Handbook of Land and Water Grabs in Africa: Foreign direct investment and food and water security 376–376.

McAuliffe, W.H.B., 2015. How did abduction get confused with inference to the best explanation? Transactions of the Charles S. Peirce Society: A Quarterly Journal in American Philosophy 51, 300–319.

Meilicke, C., Chekol, M.W., Ruffinelli, D., Stuckenschmidt, H., 2019. Anytime Bottom-Up Rule Learning for Knowledge Graph Completion. Presented at the IJCAI, pp. 3137–3143.

Miller, K.G., Browning, J.V., Schmelz, W.J., Kopp, R.E., Mountain, G.S., Wright, J.D., 2020. Cenozoic sea-level and cryospheric evolution from deep-sea geochemical and continental margin records. Science advances 6, eaaz1346–eaaz1346.
Miller, K.G., Kominz, M.A., Browning, J.V., Wright, J.D., Mountain, G.S., Katz, M.E., Sugarman, P.J., Cramer, B.S., Christie-Blick, N., Pekar, S.F., 2005. The Phanerozoic record of global sea-level change. Science (New York, N.Y.) 310, 1293–8. https://doi.org/10.1126/science.1116412
Mokrane, M., Hugo, W., Harrison, S., 2016. WDS/DSA Certification-International collaboration for a trustworthy research data infrastructure. Presented at the EGU General Assembly Conference Abstracts, pp. EPSC2016-11457.
Müller, R.D., Zahirovic, S., Williams, S.E., Cannon, J., Seton, M., Bower, D.J., Tetley, M.G., Heine, C., Le Breton, E., Liu, S., 2019. A global plate model including lithospheric deformation along major rifts and orogens since the Triassic. Tectonics 38, 1884–1907.
National Academies of Sciences, Engineering, and Medicine, 2020. A Vision for NSF Earth Sciences 2020-2030: Earth in Time. The National Academies Press, Washington, DC. https://doi.org/10.17226/25761
Nicol, A., Caruso, J., Archambault, É., 2013. Open data access policies and strategies in the European research area and beyond.
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, 1393–1409.
Peters, S.E., McClennen, M., 2016. The Paleobiology Database application programming interface. Paleobiology 42, 1–7. https://doi.org/10.1017/pab.2015.39
Pinnick, J., 2017. Exploring digital preservation requirements: A case study from the National Geoscience Data Centre (NGDC). Records Management Journal 27, 175–191. https://doi.org/10.1108/RMJ-04-2017-0009
Provost, F., Kohavi, R., 1998a. Guest Editors' Introduction: On Applied Research in Machine Learning. Machine Learning 30, 127–132. https://doi.org/10.1023/A:1007442505281
Provost, F., Kohavi, R., 1998b. Glossary of Terms. Machine Learning 30, 271–274. https://doi.org/10.1023/A:1017181826899
Raup, D.M., Sepkoski, J.J., 1982. Mass extinctions in the marine fossil record. Science 215, 1501–1503.
Ronov, A.B., Khain, V.E., Balukhovsky, A.N., Seslavinsky, K.B., 1980. Quantitative analysis of Phanerozoic sedimentation. Sedimentary Geology 25, 311–325. https://doi.org/10.1016/0037-0738(80)90067-6
Sinha, A.K., Thessen, A.E., Barnes, C.G., 2013. Geoinformatics: Toward an integrative view of Earth as a system, in: Bickford, M.E. (Ed.), The Web of Geological Sciences: Advances, Impacts, and Interactions. Geological Society of America Special Paper, pp. 591–604. https://doi.org/10.1130/2013.2500(19)
Steckel, J.C., Brecha, R.J., Jakob, M., Strefler, J., Luderer, G., 2013. Development without energy? Assessing future scenarios of energy consumption in developing countries. Ecological Economics 90, 53–67. https://doi.org/10.1016/j.ecolecon.2013.02.006
Stephenson, M.H., 2021. Affordable and Clean Energy, in: Gill, J.C., Smith, M. (Eds.), Geosciences and the Sustainable Development Goals. Springer International Publishing, Cham, pp. 159–182. https://doi.org/10.1007/978-3-030-38815-7_7

Stephenson, M.H., Cheng, Q., Wang, C., Fan, J., Oberhänsli, R., 2020. Progress towards the establishment of the IUGS Deep-time Digital Earth (DDE) programme. Episodes 43, 1057–1062.

Tierney, J.E., Poulsen, C.J., Montañez, I.P., Bhattacharya, T., Feng, R., Ford, H.L., Hönisch, B., Inglis, G.N., Petersen, S.V., Sagoo, N., 2020. Past climates inform our future. Science 370.

Torvik, R., 2009. Why do some resource-abundant countries succeed while others do not? Oxford Review of Economic Policy 25, 241–256. https://doi.org/10.1093/oxrep/grp015

Vaughan, D.J., Craig, J.R., 1978. Mineral Chemistry of Metal Sulfides. Cambridge University Press.

Wang, C., Hazen, R.M., Cheng, Q., Stephenson, M.H., Zhou, C., Fox, P., Shen, S., Oberhänsli, R., Hou, Z., Ma, X., Feng, Z., Fan, J., Ma, C., Hu, X., Luo, B., Wang, J., Schiffries, C.M., 2021. The Deep-Time Digital Earth program: data-driven discovery in geosciences. National Science Review. https://doi.org/10.1093/nsr/nwab027

Wang, C., Shen, S., Zhou, C., Hou, Z., Cheng, Q., Oberhänsli, R., Stephenson, M., 2019. Decodification of Earth Evolution in Deep-Time. Acta Geologica Sinica (English Edition) 93, 1–2. https://doi.org/10.1111/1755-6724.14225

Wilkinson, M.D., Dumontier, M., Aalbersberg, Ij.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., Van Der Lei, J., Van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B., 2016. Comment: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 1–9. https://doi.org/10.1038/sdata.2016.18

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, 5653–5658. https://doi.org/10.1073/pnas.1702297114

Zhang, X., Zhang, C., Wu, M., Lv, G., 2020. Spatiotemporal features based geographical knowledge graph construction. SCIENTIA SINICA Informationis 50, 1019–1032.

Zhao, Z., Han, S.-K., So, I.-M., 2018. Architecture of knowledge graph construction techniques. International Journal of Pure and Applied Mathematics 118, 1869–1883.

Zhou, C., Wang, H., Wang, C., Hou, Z., Zhu, Y., 2021. Prospects for the research on geoscience knowledge graph in the Big Data Era. Science China Earth Science.

Glossary:

1. **astronomical cycle**: cyclic changes in the Earth's orbital parameters such as eccentricity, axial tilt, and precession. The orbital changes lead to variation in solar radiation and Earth's climate, which is often preserved and reflected in sedimentary rocks. Also called Milankovitch cycles.
2. **artificial intelligence**: a large field that involves natural language processing, knowledge representation, automated reasoning, and so on. AI can be understood either from the perspective of internal thought processes and reasoning, or external intelligent behavior such as Turing test.
3. **association analysis**: a data analysis method that aims to find potential relationships between variables in a dataset.
4. **data-driven research paradigm**: a research philosophy that emphasizes the unprecedented importance of data capture, curation, and analysis in advancing our science in the big data era.
5. **big data**: data that are large in size and diverse in sources. Big data analysis by a variety methods, such as association analysis and machine learning, is the foundation of data-driven research paradigm in science and many other sectors.
6. **biodiversity**: the biological variety and variability of life on Earth.
7. **climate**: the average weather, or more rigorously, as the statistical description in terms of the mean and variability of relevant quantities (e.g., precipitation, temperature) over a period of time ranging from months to thousands or millions of years.
8. **chronostratigraphy framework**: a temporal framework built by correlating rock units of the same age. Chronostratigraphy framework is essential for the study of Earth history.
9. **deep-time**: time beyond Quaternary, i.e., older than 2.6 million years from present.
10. **deep learning**: a type of machine learning that uses neural networks to analyze data and identify potential patterns.
11. **DDE Data nodes**: a kind of data organization schema in DDE big data system, including national nodes, academic nodes, industry nodes, and community nodes
12. **DDE Data Alliance**: a global deep-time earth data sharing network initiated by DDE, which enabling deep-time earth data can be discovered, accessed, utilized, and recycled.
13. **FAIR principle**: the principle that data and workflow should be findable, accessible, interoperable, and reusable.
14. **feedback mechanism**: return of a fraction of the output of a system to the input as part of a cause-and-effect loop. Feedback can be positive or negative depending on the system properties and processes.

15. **greenhouse effect**: greenhouse gases such as CO2 and water vapor can absorb the infrared radiation from earth surface and therefore increase the air temperature.
16. **machine learning**: a subfield of artificial intelligence that studies the ability to improve performance based on experience.
17. **mass extinction**: a widespread and rapid decrease in the biodiversity on Earth
18. **message passing interface**: a standardized means of exchanging messages between multiple computers running a parallel program across distributed memory.
19. **mineral species**: any mineral that can be distinguished from all other minerals by current determinative methods.
20. **network analysis**: a data analysis technique that uses network theory to study complex real-world problems, such as spread of viruses, marketing, or any systems with many interacting components.
21. **plate tectonics**: a theory initially developed in the 1960s that can explain various geological processes and features very successfully by treating the lithosphere (outer rocky part of the Earth about 100-200 km thick) as a series of near-rigid blocks with nearly constant relative motion to each other.
22. **sea-level rise**: the rise of sea level compared to a reference level. Sea level rise can be either eustatic or global due to factors such as melting ice sheets, or it can be local because of factors like coastal erosion and subsidence.
23. **semantic relationship**: the relationship between the meaning of words, phrases, or sentences.
24. **trace metals**: metals that are present in small but measurable amounts in the environment.
25. **weather**: the atmospheric conditions at a particular place as regards heat, cloudiness, dryness, sunshine, wind, rainfall, etc., on a very short timescale, usually from seconds to days
26. **weathering**: the breaking down of rocks, soil, and their minerals through direct contact with the Earth's atmosphere, waters, or living things

Boxes:

Box 1: Deep time

It took 4.5 billion years for Earth to become our home today. The number is so large that many people cannot grasp it. If the entire history of the Earth was compressed into a single year, it wouldn't be until late November that we saw the first animals with hard parts. Dinosaurs died out in a mass extinction that occurred on December 26th, while human beings appeared at about 11.35 pm on New Year's Eve. Our oral tradition and written history thus misses a great majority of the epic story.

Geoscientists take up the job to translate the Earth history written in the geological record. Earth history is laid down by various physical, chemical, and biological processes. Through the study of these processes, geoscientists can partly reconstruct what happened in the past. We have a failrly good understanding of things that have taken place since Quaternary or 2.6 million years ago. However, the further we go back in time, the harder it is to say with certainty what has happened. This is because the record of the early history is usually wiped out and rewritten by later events. It is customary to call the time before Quaternary deep time because many useful records, such as tree rings, speleothems, ice cores, and corals do not exist or can hardly be used before Quaternary to reconstruct Earth history.

In DDE, one of our aim is to build high-resolution timesclae into deep time by using datasets from sedimentology, paleobiology, and plate tectonics. The timesclae will enable a comprehensive synthesis of Earth events, such as sea level changes, envioronmental changes, mass extinction, and asteroid impacts.

Box 2: Abduction and data-driven research paradigm

Abduction, also known as retroduction, is a form of reasoning that aims to explain some unexpected and curious phenomena by offering a plausible hypothesis. Abduction, along with deduction and induction, have been recognized as three different types of inference as early as by Aristotle. Their distinction has been thoroughly examined by the American philosopher Charles Sanders Peirce. Importantly, abduction generates and chooses hypotheses to test, deduction determines the entailments of a hypothesis, and induction ascertains whether the evidence accords with the hypothesis in question (McAuliffe, 2015). The three methods together form a loop through which science advances.

The enormous amount of deep-time Earth data holds the promise of revealing previously unrecognized patterns in Earth systems through the adoption of new analytic tools such as artificial intelligence and multi-dimensional analysis. The idea is that some patterns are so sophisticated and complex that they can only be identified through the analysis of large-scale datasets. In this sense, the pattern and its relevant hypothesis have been discovered abductively or are driven by big data analysis. Such a research method has been called the fourth paradigm

(Gray, 2009) and relies heavily on huge volumes of data. The findings of data-driven research are difficult, if not impossible, to be discovered by traditional research methods.

In the past, most scientific advances have come through selecting, analyzing, and visualizing just a few variables – for example, two or three different chemical elements in a mineral – and displaying the behavior of those variables in simple, two-dimensional plots (Fig. S1A). Abductive discovery (Hazen, 2014), by contrast, simultaneously considers numerous variables – a dozen or more chemical elements, for example – to discern hidden higher-dimensional patterns (Fig. S1B).

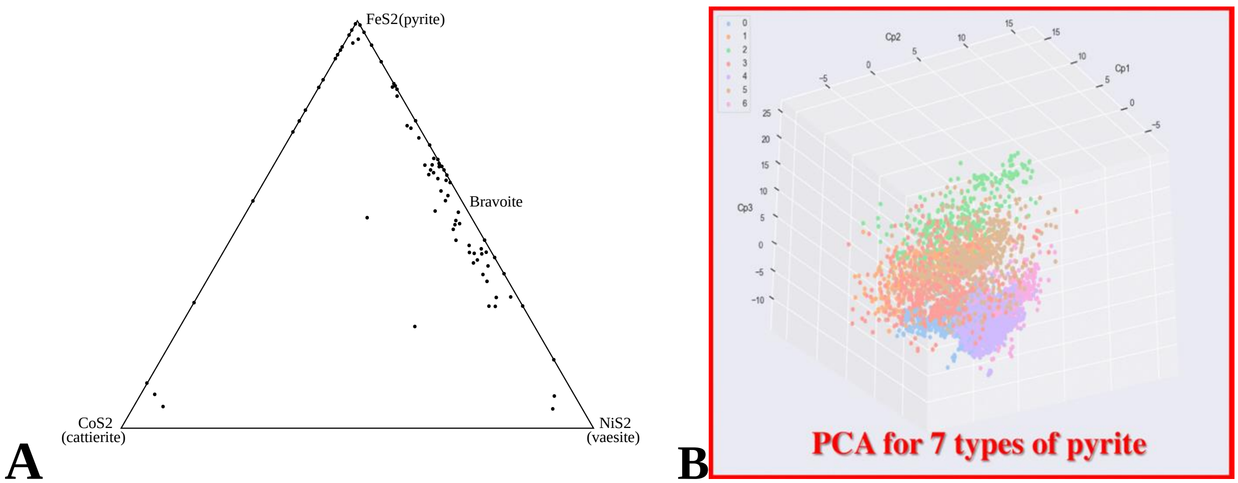

Fig. S1: (A) The traditional approach to mineral analysis involves two-dimensional plots of two or three chemical elements to discover trends in compositional data. The common mineral pyrite (FeS_2), which is often associated with valuable ore deposits, incorporates variable amounts of Fe, Ni, and Co – differences that reflect their formational environment (Bowles et al., 2011; Vaughan and Craig, 1978) (B) Cluster analysis of 12 chemical elements commonly found in trace amounts in pyrite reveals 7 distinct clusters, each representing a different ore-forming environment. Here, the 12 compositional dimensions are projected onto a three-dimensional cube. This multi-dimensional approach to pyrite analysis offers a powerful new prospecting tool (Gregory et al., 2019).

It is the hope of DDE that by integrating the isolated databases and extracting the dark data, the Earth science community can harness the power of big data and artificial intelligence, get previously unattainable insights about our planet, and transform Earth science.

Box 3: Knowledge graphs and their applications

Knowledge graphs aim to represent knowledge about the real world through a graph data model, in which the nodes are entities of interest, and the edges denote the relations among the entities. The knowledge representation is achieved mostly by a subject-predicate-object triple stored in the graph. For instance, the fact that "sandstones are a type of sedimentary rocks" can be represented as <sandstone, isTypeof, sedimentary_rock>, where *sandstone* is the subject, *isTypeof* is the predicate, and *sedimentary_rock* is the object. Subjects and objects are represented as nodes in the knowledge graph, while the predicate is typically represented as the edge. It is through the interlinking of entities that knowledge graphs establish a meaningful and machine-readable representation of the real world.

Knowledge graphs have broad applications ranging from information extraction, natural language processing, question answering, automated reasoning to personalized recommendations due to their machine-readable nature and rich semantic relationships. Prominent examples powered by knowledge graphs include Apple's Siri service, Google's Knowledge Graph, and IBM's Watson. In Earth science, previous endeavors such as the Semantic Web for Earth and Environment Terminology (SWEET) and Extraction of Dark Data (xDD) have shown great potential of knowledge graphs.

In DDE, we aim to establish a whole-domain knowledge graph through expert crowdsourcing and machine learning to facilitate our data integration and analysis processes. In the end, we hope to build an intelligent research assistant that can help Earth scientists reach their greatest potential by providing services to every step in the research process.

Appendix:

Data center name	**Database theme**	URL
OneGeology	Geological map	http://www.onegeology.org/home.html
Geochemical Databases for the Earth (EarthChem)	Petrology (mainly magmatic rock)	https://www.earthchem.org
Geochemistry of Rocks of the Oceans and Continents (GEOROC)	Magmatic rock	http://georoc.mpch-mainz.gwdg.de/georoc/
Geochron	Geochronology	https://www.geochron.org
DataView	International Chronology and Isotope Database	http://sil.usask.ca/databases.htm
Neptune Sandbox Berlin (NSB)	Paleontology	http://www.nsb-mfn-berlin.de/
Archives of Digital Morph (ADMorph)	Paleontology	http://www.admorph.org/
PALEOMAP	Paleogeography	http://www.scotese.com
Neotoma	Paleontology	http://www.neotomadb.org
TimeScale Creator	Stratigraphy	https://timescalecreator.org/index/index.php
Chinese Vertebrate Paleontology and Paleoanthropology Database (VPPDB)	Ancient vertebrates	http://www.deepbone.org
Aster images	Remote sensing	https://earthexplorer.usgs.gov/
National Geochronological Data Base (NGDB)	Geochronology	https://mrdata.usgs.gov/geochron/map-us.html#home
Paleobiology Database (PBDB)	Paleontology	http://fossilworks.org/
British Geological Survey (BGS)	Paleontology	https://www.bgs.ac.uk/palaeosaurus/home.cfm http://www.3d-fossils.ac.uk/
Global Biodiversity Information Facility (GBIF)	Paleontology	https://www.gbif.org
Japan Paleobiology Database (jPaleoDB)	Paleontology	http://jpaleodb.org/
Mikrotax	Paleontology	http://www.mikrotax.org/
Fossil Record Electronic Database (FRED)	Paleontology	https://fred.org.nz

Neogene Mammal Mapping Portal (NEOMAP)	Ancient vertebrates	https://ucmp.berkeley.edu/neomap/
FAUNMAP	Ancient vertebrates	https://ucmp.berkeley.edu/faunmap/index.html
MioMap	Ancient vertebrates	https://ucmp.berkeley.edu/miomap/
New and Old Worlds: Database of Fossil Mammals (NOW)	Ancient vertebrates	http://www.helsinki.fi/science/now/
Canadian Database of Geochemical Surveys	Geochemistry	https://geochem.nrcan.gc.ca/cdogs/content/main/home_en.htm
Landsat 8 Images	Remote sensing	https://glovis.usgs.gov/
Ziyuan-1 02D	Remote sensing	http://www.sasclouds.com/chinese/home
Gaofen-5	Remote sensing	http://www.sasclouds.com/chinese/satellite/chinese/gf5
Mineral Resources Data System	Mineral resources	https://mrdata.usgs.gov/mrds/
SFB 267	Geochronology	http://www.cms.fuberlin.de/sfb/sfb267/results/data_catalogue/central_andean_data/geochemical_data.html
NGU GEOCHRON	Geochronology	http://geo.ngu.no/kart/geokronologi_mobil/?lang=eng
Petlab Database	Geochronology	https://pet.gns.cri.nz
Greenland U-Pb Geochronology Database	Geochronology	http://www.greenmin.gl
Open Cosmogenic Isotope and Luminescence Database(OCTOPUS)	Geochronology	https://earth.uow.edu.au
Utah Geochronology Database	Geochronology	https://geology.utah.gov/apps/geochron/
Yukon Geochronology database, YGD	Geochronology	http://data.geology.gov.yk.ca/Com-pilation/22
Geoscience Australia's Geochron Delivery system (GAGDS)	Geochronology	http://www.ga.gov.au/geochron-sapub-web/geochronology/shrimp/search.Htm

Canadian Geochronology Knowledgebase（CGKB）	Geochronology	https://www.nrcan.gc.ca/earth-sciences/geography/atlas-canada/canadian-geo-chronology-knowledgebase/18211#cgkb
PALEOMAGIA	Geochronology	https://h175.it.helsinki.fi/database/form.php
Geobiodiversity Database(GBDB)	Stratigraphy	http://www.geobiodiversity.com/
Geological Survey of Japan (GSJ)	Paleontology	https://www.gsj.jp/
Fossiilid.info	Paleontology	https://fossiilid.info/
Biological library (BioLib)	Paleontology	https://www.biolib.cz/
ZooBank	Paleontology	http://www.zoobank.org
The Bibliography of Fossil Vertebrates (BFV Online)	Paleontology	http://www.bfvol.org/
Ancient Human Occupation of Britain（AHOB）	Ancient vertebrates	http://www.ahobproject.org/
Morphbank	Ancient vertebrates	https://www.morphbank.net/
MorphoBank	Ancient vertebrates	https://morphobank.org/
World Stress Map (WSM)	Plate tectonics	http://www.world-stress-map.org
Marine Geoscience Data System (MGDS)	Plate tectonics	www.marine-geo.org/about/overview.php
MetPetDB	Metamorphic petrology	http://metpetdb.com/
Macrostrat	Sedimentology	https://macrostrat.org
Ava Clastics	Sedimentology	https://www.pds.group/ava-clastics
Sedimentary Analogs Database and Research Consortium (SAnD)	Sedimentology	https://geology.mines.edu/research/sand/
Geomagnetism	Geomagnetism	http://www.geomag.us/
GEOMAGIA50	Geomagnetism	https://www.gfz-potsdam.de/en/section/geomagnetism/data-products-services/geomagia50/
International Association of Geomagnetism and Aeronomy (IAGA)	Geomagnetism	http://www.iaga-aiga.org/

SedDB	Geochemistry	http://www.earthchem.org/seddb
China Seismic Array Data Management Center	Geophysics	http://www.chinarraydmc.cn/
EarthScope	Geophysics	https://www.earthscope.org/
International Geothermal Association (IGA)	Geothermal	https://www.geothermal-energy.org
Geothermal Resources Council (GRC)	Geothermal	https://www.geothermal.org
HeatFlow.org	Geothermal	http://heatflow.org/
National Geothermal Data System (NGDS)	Geothermal	http://geothermaldata.org/
Geothermal District Heating (GEODH)	Geothermal	http://geodh.eu/
China Heat Flow Database	Geothermal	http://chfdb.xyz/
General Bathymetric Chart of the Oceans (GEBCO)	Geothermal	www.gebco.net
National Ceters for Environmental Information (NCEI)	Geomagnetism	https://www.ngdc.noaa.gov/
GeoCloud	Geology	http://geocloud.cgs.gov.cn/#/portal/home
OpenGeoscience	Geological map	https://www.bgs.ac.uk/geological-data/opengeoscience/
National Geologic Map Databaseroject (NGMDB)	Geological map	https://ngmdb.usgs.gov/
VertNet	Ancient vertebrates	http://www.vertnet.org/index.html
Sepkoski's Online Genus Database	Paleontology	http://strata.geology.wisc.edu/jack/
National Infrastructure of Mineral, Rock and Fossil Resources for Science and Technology	Fossil specimen	http://www.nimrf.cugb.edu.cn/
mindat.org	Mineralogy	https://www.mindat.org/
SinoProbe Center	Deep geology	www.sinoprobe.org
PANGAEA	Petrology, mineralogy, environmental science	https://www.pangaea.de

Integrated Ocean Drilling Program (IODP)	Petrology	http://www.iodp.org/
Petrological Database of the Ocean Floo (PetDB)	Petrology	https://search.earthchem.org/
J-CORES	Drill core	http://sio7.jamstec.go.jp/contents/
Data and Sample Research System for Whole Cruise Information in JAMSTEC (DARWIN)	Marine Geology	http://www.godac.jamstec.go.jp/
DATAONE	Earth Science	https://www.dataone.org/
Earth System Science Data (ESSD)	Earth Science	https://www.earth-system-science-data.net/index.html
GeoGratis	Geography	http://geogratis.gc.ca
Earth Cube (EC)	Earth Science	https://www.earthcube.org/