Received: 26 July 2022 | Revised: 16 November 2022 | Accepted: 6 January 2023

DOI: 10.1002/gdj3.186

DATA SERVICES ARTICLE

Geoscience Data Journal RMetS WILEY

GeoDeepShovel: A platform for building scientific database from geoscience literature with AI assistance

Shao Zhang[1] | **Hui Xu**[1] | **Yuting Jia**[1] | **Ying Wen**[1] | **Dakuo Wang**[2] | **Luoyi Fu**[1] | **Xinbing Wang**[1] | **Chenghu Zhou**[3]

[1]Shanghai Jiao Tong University, Shanghai, China

[2]IBM Research, Cambridge, Massachusetts, USA

[3]Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China

Correspondence
Ying Wen, School of Electronic Information Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China.
Email: ying.wen@sjtu.edu.cn

Funding information
National Natural Science Foundation of China, Grant/Award Number: 42050105 and 62106141; Shanghai Sailing Program, Grant/Award Number: 21YF1421900

Abstract

With the rapid development of big data science, the research paradigm in the field of geosciences has also begun to shift to big data-driven scientific discovery. Researchers need to read a huge amount of literature to locate, extract and aggregate relevant results and data that are published and stored in PDF format for building a scientific database to support the big data-driven discovery. In this paper, based on the findings of a study about how geoscientists annotate literature and extract and aggregate data, we proposed GeoDeepShovel, a publicly available AI-assisted data extraction system to support their needs. GeoDeepShovel leverages state-of-the-art neural network models to support researcher(s) easily and accurately annotate papers (in the PDF format) and extract data from tables, figures, maps, etc., in a human–AI collaboration manner. As a part of the Deep-Time Digital Earth (DDE) program, GeoDeepShovel has been deployed for 8 months, and there are already 400 users from 44 geoscience research teams within the DDE program using it to construct scientific databases on a daily basis, and more than 240 projects and 50,000 documents have been processed for building scientific databases.

KEYWORDS

artificial intelligence, big data-driven discovery, data extraction, scientific database, human-computer interaction

1 | INTRODUCTION

The rapid development of big data-related technologies makes big data-driven scientific research increasingly important in geosciences. Geoscience research often relies on a large amount of exploration data. The shifting of the data-driven research paradigm raises new requirements for researchers to build *scientific databases* (Hoeppe, 2021) in Geoscience (Bergen et al., 2019). *Scientific database* is a collection of structured and verified research results that consist of various numeric, word-oriented or image-organized data, which plays a central role in data-driven research (National Research Council, Division on Engineering and Physical Sciences, Commission on Physical Sciences, Mathematics, and Applications, Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest et al., 2000). The collection and organization of scientific data is a critical step in the research process. Scientific data as an infrastructure, the source of data, accuracy, depth, breadth of data and other perspectives jointly affect the research progress. In the field of geoscience, a single research team is limited by region and time and often cannot rely on itself to complete a large amount of original data collection, which also makes the reuse, sharing and disclosure of scientific data an essential issue in geoscience research. The FAIR (findable, accessible, interoperable and reusable) Guiding Principles (Wilkinson et al., 2016) make the construction and management of current scientific data have a common goal. And the Deep-Time Digital Earth (DDE) program (Oberhänsli, 2020), which is a data-driven discovery program in geoscience with a goal of aggregating the geoscience data and facilitating data-driven discovery for understanding Earth's evolution (Wang et al., 2021), also puts forward the vision of co-construction and sharing of geoscience data.

However, the DDE program finds that current difficulties of data-driven discovery include the lack of digitization, and databases do not adhere to FAIR principles (Wang et al., 2021). There are still lots of research data scattered in the past literature which needs to be collected and sorted so that they can be shared and reused. Therefore, the collection and collation of data in the past literature is a very important and arduous task, and many research groups are still committed to it. Geoscientists often review a large amount of published literature to obtain enough high-quality data (McMahon & Davies, 2018; Puetz, 2018; Puetz et al., 2018) (generally PDF documents), from which they locate and extract valuable data (e.g. tables, figures, maps, etc.) to construct the scientific databases. Many influential studies have also been built on such data collection efforts, such as Fan et al. (2020), Dirzo et al. (2014) and Tucker et al. (2018).

The current literature is often disseminated in the form of PDF, and these data are stored in unstructured form, including pictures, tables and texts. The typical way is to manually read the literature to extract these data and organize and write them into the scientific database (see Figure 1). The traditional manual extraction process has a low degree of automation and consumes a lot of human resources and material resources, which largely hinders small teams from conducting research related to big data. Although some larger research teams may have more workforce, without a well-designed collaborative platform, they still need to spend lots of effort to process a sufficient amount of literature and extract enough data for the scientific database (Renaudie et al., 2020). Because of these challenges, constructing a scientific database using the data extracted from a large number of papers often takes several years with a large workforce, which is a massive obstacle to the advancement of research.

In this work, we focus on the need to construct scientific databases in geoscience, which can help geoscientists to discover unknown phenomena and novel insights into Earth (Dirzo et al., 2014; Fan et al., 2020; Tucker et al., 2018). Scientific data research works often focus on the analysis and research of datasets but ignore the dataset construction process and the difficulties researchers encounter in this process. As a part of the DDE program, our research is committed to building an AI-assisted platform to help geoscience research teams complete data extraction, integration and storage in one-stop, and builds a scientific database to make the data conform to the FAIR principle.

In the past, many researchers in geoscience tried to build an integrated platform from data collection and storage to data analysis, which promoted the sharing of geoscience databases and big data research. Chronos (Cervato et al., 2005) is a community facility addressing the needs of geoinformatics and providing simultaneous and seamless integration of hosted and federated databases with analytical and visualization tools. Paleostrat (Snyder et al., 2008) is designed as an infrastructure platform for Geoscience researchers and teachers, which serves the community by enhancing the research and education process. However, they mainly focus on the design of the schema for the data utilized instead of trying to make a user-friendly interface for geoscientists to easily extract desired information. It leads to the result that these platforms can hardly generate enough data to support a healthy life cycle. GeoDeepDive (Zhang et al., 2013) is a widely used toolkit that adopts natural language processing (NLP) technology to process and analyse the literature end-to-end. However, due to the complete dependence on the end-to-end extraction method (Govindaraju et al., 2013), the lack of labelled data has introduced the problem of insufficient data accuracy, which leads to a result that the data extracted by GeoDeepDive cannot play a great role in the research that requires accurate data for modelling and analysis (Sun et al., 2022). Moreover, these methods can only cover the text part (Ashktorab et al., 2021; Desmond et al., 2021; Niu et al., 2012), use NLP to analyse the tables with the content (Govindaraju et al., 2013) and cannot process data from tables and pictures, which leads to damage to the integrity of the data, which is also detrimental to research. Besides, researchers would have to pay extra effort to label and clean the training data to make the end-to-end AI model work, which may take even more time than manually extracting data without using an AI.

In this paper, we argue that instead of designing a fully automated end-to-end solution, a human–AI collaborative and interactive system may be the solution to address these problems. Geoscientists can perform the data extraction activity as they used to, and AI can train itself with these user-labelled data and then make suggestions

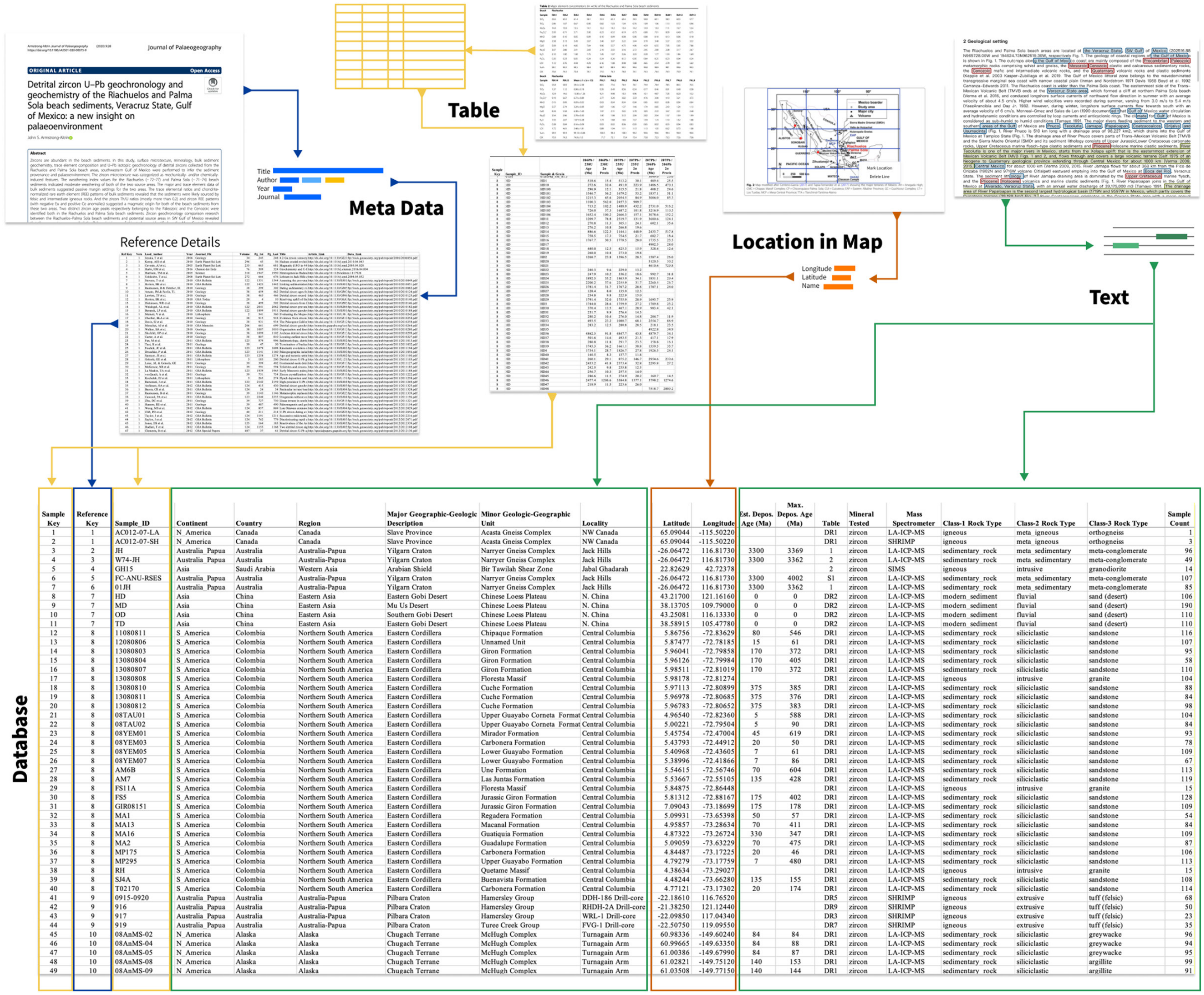


FIGURE 1 A workflow's example for extracting data from a literature and save to database. The researchers extract the corresponding values of the attributes of the samples from different parts of the article and fill them in the dataset. (Figures and table are from (Armstrong-Altrin, 2020). The database is from (Puetz, 2018). This flowchart is just show the workflow but does not show the real data.).

to the user in the future. Together, the human–AI team can accurately extract data at a much lower cost. Considering the work on extracting and integrating data from past literature and building databases in data-driven geoscience discovery, we designed GeoDeepShovel, an artificial intelligence-based collaborative data extraction platform, to assist geoscience research teams in data extraction, data aggregation and scientific database construction. GeoDeepShovel provides a user-friendly interface and experience design following the human–AI interaction design guidelines (Amershi et al., 2019) so that users even without any AI backgrounds can also work comfortably with it. In such a human–AI collaboration process, we use human participation to ensure the precision and accuracy of the data. The assistance of AI can greatly reduce the manual workload of humans. Researchers only need to judge whether the data are correct and make small corrections to extract and import the data into the database. Further, GeoDeepShovel can extract data in pictures and tables, which greatly increases the diversity of data, which can make the extracted research data more complete and facilitate subsequent analysis and research. GeoDeepShovel has been deployed for 8 months and there are already 400 users from 44 geoscientist teams within the DDE program using it on a daily basis. We have cooperated with the OneSediment team in the DDE program and 26 thematic databases in OneSediment have used GeoDeepShovel for data extraction currently.

We first analyse the data content of current scientific databases in the earth sciences, describe the distribution and extraction process of these data in the article and discuss the difficulties existing in the current process

(Section 2). We provide an overview of the system we have built in Section 3. We detail how researchers use our system to extract data scattered across literature texts, images and tables with the aid of artificial intelligence while helping our model gather training data. After data extraction, researchers can integrate the data extracted from multiple documents into one file according to the storage method of the database, which is convenient for data storage. We conclude by discussing how we will continue to advance this work in the future (Section 4). Our work combines professional knowledge from multiple disciplines such as artificial intelligence, human–computer interaction, data management, software design and earth science research, and has conducted sufficient user research and verification, hoping to provide data-driven scientific research in geoscience a guide of data extraction working practices.

2 | CHALLENGES TO SUPPORT DATA EXTRACTION WORKFLOW FOR SCIENTIFIC DISCOVERY

2.1 | Scientific database in geoscience and current workflow of construction

Scientific database is a collection of structured and verified research results that consists of various numeric, word-oriented or image-organized data, which plays a central role in data-driven research (National Research Council, Division on Engineering and Physical Sciences, Commission on Physical Sciences, Mathematics, and Applications, Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest et al., 2000). We take "A relational database of global U–Pb ages"(Puetz, 2018) as an example to explain the composition of a scientific database in geoscience research. This database contains 700,598 records of global U–Pb ages. The data are restructured and made available as a relational database. The database is available in two formats – a Microsoft® Excel™ version with only the basic data and a Microsoft® Access™ version with the basic data plus graphic functionality.

In the Microsoft® Excel™ version (the screenshot shown in Figure 1 Dataset part), there are six sheets including Reference Details, Sample Details, Data, Rock Type Lists, Other Lists and Summary. The Reference Details sheet includes the reference articles' meta-information and the reference Key. Sample Details sheet shows each sample from different articles and the attributes of the samples such as geographic location, latitude and longitude information, lithology, original sample ID, etc. Moreover, the Data sheet is the main data containing the U–Pb ages of each sample. These three sheets are related by the reference Key and the sample Key. Rock Type Lists, Other Lists and Summary are some information about the attributes and the database.

In the Microsoft® Access™ version, more detailed data are provided expect the information in the Microsoft® Excel™ version. There are 20 tables in the database as shown in Figure 2.

From this example, we can find that the main information the dataset/database provided is the source of the data and various geologically relevant attributes of the samples. Compared with the Microsoft® Excel™ version, we can find that the Microsoft® Access™ version provides more detailed information, including the journal list of the articles and the statistics and summaries of some items in the dataset. However, the core data about the sample and its description are consistent. This indicates that the data in the Microsoft® Excel™ version are completely available for the study, and the additional details in the Microsoft® Access™ version are used for some statistics and corroboration of the data. This is consistent with the findings of our interviews

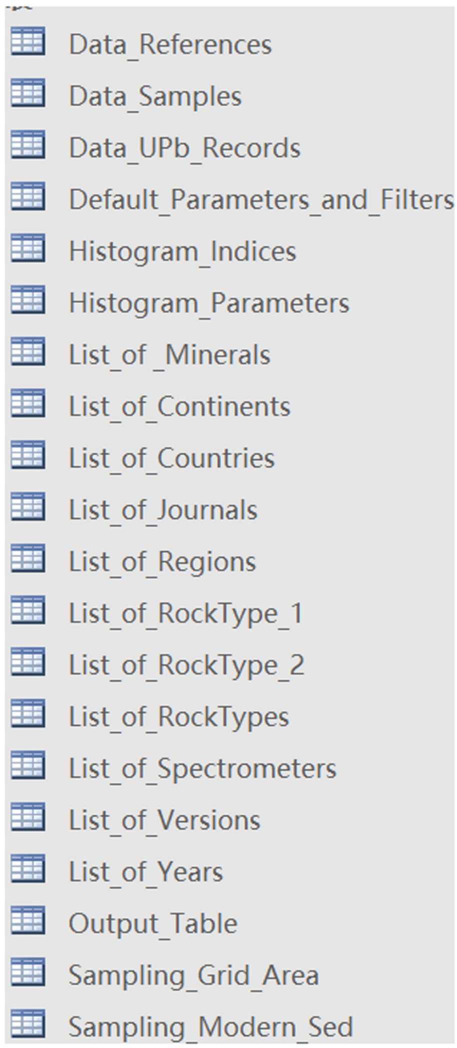


FIGURE 2 The tables' list of the MicrosoftR Access™ version database.

and studies with many research teams in the DDE program. Geoscience research teams typically use Microsoft® Excel™ to store data briefly and use a Microsoft® Excel™ version for publication. Furthermore, the data storage form is generally Microsoft® Excel™ (Brand et al., 2015; Puetz et al., 2018) and some commonly used database formats, such as Microsoft® Access™(Puetz, 2018) and MySQL.

Although researchers from different sub-fields of geoscience may study different research questions and focus on different data, their research processes and workflows are basically the same. Figure 1 shows the typical workflow of the database construction for big data discovery in common geoscience research. Based on the practice in DDE program, we conclude the task in this workflow (see Figure 3). These tasks are grouped and detailed by the main workflow steps in the following list:

- **T1-Problem Definition:** Define the research problem and the structure of the scientific database that needs to be built,
- **T2-Search:** Search for the paper that may contain data about the research problem,
- **T3-Scan:** Quickly scan the article to find data that is needed,
- **T4-Meta-Information Extraction:** Record the literature's meta-information for tracking the data,
- **T5-Detail Data Extraction:** Extract data from different parts of the literature,
  - **Data extraction from the table:** Get the data in the table and fill in the Microsoft® Excel™ file prepared in an advance cell by cell,
  - **Information extraction from the text:** Search the full text with keywords to locate the data, fill in the data in the Microsoft® Excel™ file after finding it and repeat until all the data are found,
  - **Information extraction from the figure:** Restore the corresponding information from the images, especially obtaining the latitude and longitude of a marked point from the map,
- **T6-Proofreading:** Check and proofread the data to ensure the data are accurate,
- **T7-Data Integration:** Integrate the data extracted from each paper (usually stored in a bunch of Microsoft® Excel™ files) into the final dataset (Figure 3).

2.2 | Challenges from multi-modal data in geoscience literature

The construction of a scientific database for big data research in geosciences requires data from four different parts of the literature: literature metadata, text, tabular data and map.

When extracting data, the most basic and essential point is to record the source of the data, which is the Metadata of the literature. The purpose of recording article meta-information is to make each data entry traceable. Therefore, this meta-information will be integrated into each data entry extracted from the current article as part of the attributes. The metadata includes title, author(s), published year, keywords, publisher, ISSN, volume, issue, DOI, language and other additional information. Metadata usually appear on the first page of an article, with some formatting differences due to different publishers and journal page layouts (as shown in Figure 4). The different formats bring difficulty to extract the metadata manually. It takes a long time to extract literature metadata, requiring multiple step-by-step replications. Typically, researchers use academic search engines such as Google Scholar and the Web of Science to obtain the metadata (McMahon & Davies, 2018), but this requires additional searching and sorting, which still results in much time being wasted on mechanical tasks.

Tables are the best way to carry large amounts of data in scientific literature. In the geoscience literature, authors often use tables to present measurements and chemical analyses of the samples. There are specific differences in each author's writing habits and research process, which makes the styles of these tables also vary (as shown in Figure 5).

Researchers usually use some OCR tools with graphical user interfaces to process PDF documents to make them editable and then manually copy-paste the data they

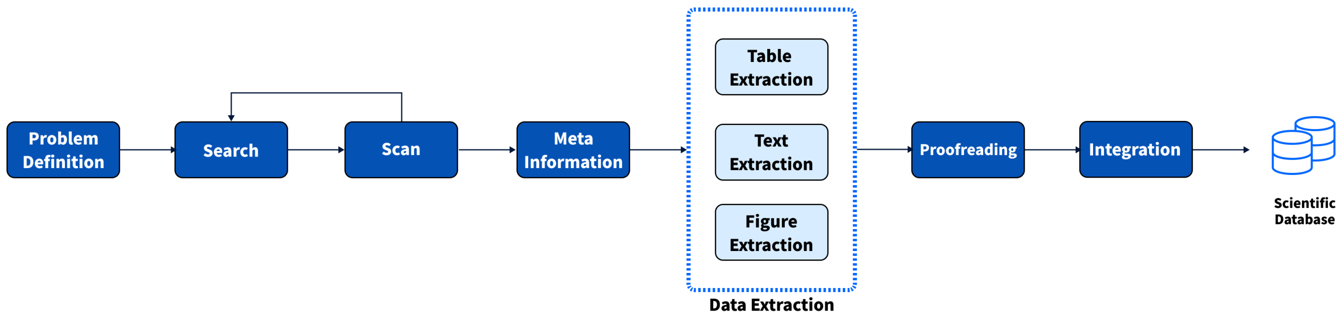


FIGURE 3 The team's workflow of collecting and building a database from geoscience literature.

find into Microsoft® Excel™. This method will bring a huge workload and is prone to errors when the number of rows and columns in the table is large. The other type is that the author publishes the data in the form of an appendix. However, the file formats of the appendix are very diverse, including but not limited to *.doc, *.xlsx, *.pptx and *.csv because there is no unified specification, which also brings more incredible difficulty to manual extraction. As a result, geoscientific researchers unfamiliar with related programming techniques often choose the most primitive way of copying and pasting one by one for extraction, which makes the workload extremely large.

As the main part of the article, the text contains the research object of the article and its related attributes, including but not limited to the name, lithology, geological age and geographical location of the research object (as shown in Figure 6). This information is an essential part of the scientific dataset. However, the text part of the article is unstructured, and due to the different writing styles of researchers and the different organization of research work, the distribution of this information is not regular to follow. At the same time, since there may be multiple research object subjects in an article, the distribution of this information may be more complicated (such as using "respectively" sentences to describe the value of the same attribute of two objects). In manual extraction, it is often necessary to rely on the researcher's professional domain knowledge and experience from reading literature to search for the needed information. Because the information eventually needs to be organized into a database, researchers often spend much time linking the information together. For example, a description such as "sample A was collected at location B" needs to be defined and stored as a triple "sample A - collection location - location B." This process of linking information can also be extremely difficult without the support of tools.


(a)

Armstrong-Altrin *Journal of Palaeogeography* (2020) 9:28
https://doi.org/10.1186/s42501-020-00075-9

Journal of Palaeogeography

ORIGINAL ARTICLE Open Access

Detrital zircon U–Pb geochronology and geochemistry of the Riachuelos and Palma Sola beach sediments, Veracruz State, Gulf of Mexico: a new insight on palaeoenvironment

John S. Armstrong-Altrin

Abstract

Zircons are abundant in the beach sediments. In this study, surface microtexture, mineralogy, bulk sediment geochemistry, trace element composition and U–Pb isotopic geochronology of detrital zircons collected from the Riachuelos and Palma Sola beach areas, southwestern Gulf of Mexico were performed to infer the sediment provenance and palaeoenvironment. The zircon microtexture was categorized as mechanically- and/or chemically-induced features. The weathering index values for the Riachuelos (~ 72–77) and Palma Sola (~ 71–74) beach sediments indicated moderate weathering of both of the two source areas. The major and trace element data of bulk sediments suggested passive margin settings for the two areas. The trace elemental ratios and chondrite-normalized rare earth element (REE) patterns of bulk sediments revealed that the sediments were likely sourced by felsic and intermediate igneous rocks. And the zircon Th/U ratios (mostly more than 0.2) and zircon REE patterns (with negative Eu and positive Ce anomalies) suggested a magmatic origin for both of the beach sediments from these two areas. Two distinct zircon age peaks respectively belonging to the Paleozoic and the Cenozoic were identified both in the Riachuelos and Palma Sola beach sediments. Zircon geochronology comparison research between the Riachuelos–Palma Sola beach sediments and potential source areas in SW Gulf of Mexico revealed that the source terrane supplied the Paleozoic zircons of this study was identified as the Mesa Central Province (MCP), and the Cenozoic zircons were transported from the nearby Eastern Alkaline Province (EAP). Moreover, although the Precambrian zircons were very few in the studied sediments, their geochronology and geochemistry results still could infer that they were contributed by the source terranes of Grenvillian igneous suites in the Oaxaca and the Chiapas Massif Complexes.

Keywords: Detrital zircon, Beach sediment, U–Pb dating, Zircon grain morphology, Microtexture, Mineralogy, Geochemistry, Geochronology, Gulf of Mexico

Correspondence: armstrong@cmarl.unam.mx; john_arms@yahoo.com
Unidad de Procesos Oceánicos y Costeros, Instituto de Ciencias del Mar y Limnología, Universidad Nacional Autónoma de México, Ciudad Universitaria, 04510 Ciudad de México, Mexico

Springer Open

(b)

RESEARCH

PALEONTOLOGY

A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity

Jun-xuan Fan[1,2], Shu-zhong Shen[1,2,3*], Douglas H. Erwin[4,5], Peter M. Sadler[6], Norman MacLeod[1], Qiu-ming Cheng[7], Xu-dong Hou[1], Jiao Yang[1], Xiang-dong Wang[1], Yue Wang[2], Hua Zhang[2], Xu Chen[2], Guo-xiang Li[2], Yi-chun Zhang[2], Yu-kun Shi[1], Dong-xun Yuan[2], Qing Chen[2], Lin-na Zhang[2], Chao Li[2], Ying-ying Zhao[1]

One great challenge in understanding the history of life is resolving the influence of environmental change on biodiversity. Simulated annealing and genetic algorithms were used to synthesize data from 11,000 marine fossil species, collected from more than 3000 stratigraphic sections, to generate a new Cambrian to Triassic biodiversity curve with an imputed temporal resolution of 26 ± 14.9 thousand years. This increased resolution clarifies the timing of known diversification and extinction events. Comparative analysis suggests that partial pressure of carbon dioxide (P_{CO_2}) is the only environmental factor that seems to display a secular pattern similar to that of biodiversity, but this similarity was not confirmed when autocorrelation within that time series was analyzed by detrending. These results demonstrate that fossil data can provide the temporal and taxonomic resolutions necessary to test (paleo)biological hypotheses at a level of detail approaching those of long-term ecological analyses.

Fan et al., Science 367, 272–277 (2020) 17 January 2020 1 of 6

FIGURE 4 Different layouts of the articles' first pages form of an appendix. However, the 148 file formats of the appendix are very diverse. (a) First page layout of an article from Journal of Palaeogeography (Armstrong-Altrin, 2020). (b) First page layout of an article from Science (Fan et al., 2020).

FIGURE 5 Different Table Styles of Articles. (a) A table from Detrital zircon U–Pb geochronology and geochemistry of the Riachuelos and Palma Sola beach sediments, Veracruz State, Gulf of Mexico: a new insight on palaeoenvironment (Armstrong-Altrin, 2020). (b) A table from Global database of diffuse riverine nitrogen and phosphorus loads and yields (McDowell et al., 2021).

(a)

Table 1 Microtextures of mechanical and chemical features identified on the zircon grain surfaces in the Riachuelos and Palma Sola beach sediments

Microtexture	Zircon grain		Palaeoenvironment[a]
	Riachuelos	Palma Sola	
Mechanically-induced feature			
Abraded edge (abe)	X	XXX	Aeolian, saltation, collision
Dual striated zircon (dsz)	X	XX	Saltation, collision, short transport
Euhedral zircon with one side broken edge (bez)	XX	X	Aeolian, saltation, collision, short transport, storm record
Crescentic gouge (crg)	X		Near shore, wave action
Arc-shaped step (as) and Linear step (ls)	X	XX	High-energy collision, aeolian, littoral zone, glacial zone
Bulbous edge (ble)	XX		Aeolian, saltation, fluvial, dune
Reworked conchoidal fracture (rcf)	X	XX	High-energy collision, aeolian, littoral zone, nearshore subaqueous
Collision fracture (cf)		XXX	High-energy collision, aeolian, littoral zone
Meandering ridge (mr)		X	Aeolian, littoral dune, subaqueous
V-shaped percussion crack (vs)		X	High-energy collision, gouging, littoral zone, deltaic, subaqueous, surf zone
Straight groove (sgr)	X		Littoral zone, wave action, saltation
Chemically-induced feature			
Solution and precipitation feature (s/p)	XX	XX	Diagenetic environment, high in contaminated sea water (alkaline fluid)
Circular solution pit (csp)	XX	X	Intertidal zone, diagenetic, percolation of sea water
Grain cavities (gc)			Diagenetic, percolation of sea water
Delamination (dl)	X	X	Collision/diagenetic
Silica pellicle (sp)	X	X	Starting stage of in-situ diagenetic, nearshore
Adhered particle appears to be silica globule (ads)		X	In-situ diagenetic, silica saturated, low-energy
Silica flower (sf) and crystal overgrowth		X	Advanced stage of diagenetic environment, silica oversaturated nearshore, low-energy
Adhered particle (ad)[b]	XXX	XXX	Diagenetic, littoral, low-energy

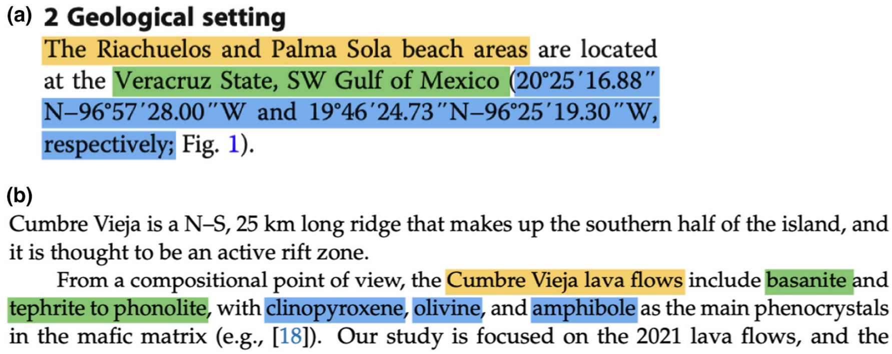
XXX means Abundant; XX means Common; X means Present; [a] stands for citations after Mahaney (2002), Madhavaraju et al. (2009), Mahaney et al. (2012), Armstrong-Altrin and Natalhy-Pineda (2014), Vos et al. (2014), Hossain et al. (2020), and Mohammad et al. (2020); [b] Adhered particles are defined as mechanical/chemical origin in some studies (e.g. Madhavaraju et al. 2009; Vos et al. 2014). Referring Figs. 6 and 7 for SEM images

(b)

TABLE 1 Data sources and richness at several steps used to calculate the load and yields of phosphorus and nitrogen forms

Database	N or P fraction	Step 1: Number of sites (data records) after checking	Step 3: Number of sites (data records) after filtering and harmonization	Steps 5 and 6: Number of catchments with predictor variables contributing to global yield models	Source of data
GEMStat	DRP	1,392 (59,065)	111 (12,310)	107	United Nations Environment Programme (2018)
	TP	2,268 (80,120)	640 (40,897)	254	
	NO_x-N	605 (35,870)	136 (12,863)	107	
	TN	2,476 (119,526)	744 (55,181)	76	
GLORICH	DRP	818 (137,461)	579 (104,590)	486	Hartmann *et al*. (2014)
	TP	770 (110,260)	481 (74,020)	481	
	NO_x-N	717 (152,886)	517 (99,944)	378	
	TN	517 (68,167)	316 (39,892)	158	
Murray-Darling	DRP	25 (34,642)	25 (18,325)	10[a]	Biswas and Mosley (2018)
	TP	25 (35,889)	25 (18,611)	10	
	NO_x-N	25 (34,454)	24 (17,733)	9	
	TN	30 (45,852)	29 (22,100)	10	
NZWQ	DRP	723 (74,573)	449 (31,121)	438	McDowell *et al*. (2013); Larned *et al*. (2016)
	TP	723 (74,571)	449 (31,121)	311	
	NO_x-N	723 (74,573)	449 (31,121)	86	
	TN	723 (74,573)	449 (31,121)	26	

[a]Only used as part of the technical validation.


(a) 2 Geological setting
The Riachuelos and Palma Sola beach areas are located at the Veracruz State, SW Gulf of Mexico (20°25′16.88″ N–96°57′28.00″W and 19°46′24.73″N–96°25′19.30″W, respectively; Fig. 1).

(b)
Cumbre Vieja is a N–S, 25 km long ridge that makes up the southern half of the island, and it is thought to be an active rift zone.
From a compositional point of view, the Cumbre Vieja lava flows include basanite and tephrite to phonolite, with clinopyroxene, olivine, and amphibole as the main phenocrystals in the mafic matrix (e.g., [18]). Our study is focused on the 2021 lava flows, and the

FIGURE 6 Examples of research object descriptions in the text. The pictures are from the screenshots of the article PDFs. In different articles, the authors use different sentences to describe similar information. Differences in article publishers and templates further lead to different forms of text presentation, which brings challenges to digital processing. (a) The location description of the research area from Detrital zircon U–Pb geochronology and geochemistry of the Riachuelos and Palma Sola beach sediments, Veracruz State, Gulf of Mexico: a new insight on palaeoenvironment (Armstrong-Altrin, 2020). The yellow highlight is the area name, the green highlight is the location and the blue highlight is the latitude and longitude. (b) The lithology description of the research object from Rock Magnetism of Lapilli and Lava Flows from Cumbre Vieja Volcano, 2021 Eruption (La Palma, Canary Islands): Initial Reports (Parés et al., 2022). The yellow highlight is the research object name, the green highlight is main lithology and the blue highlight is the detailed lithology.

Geographical location information, as fundamental data in geoscience, often appears in articles in the form of maps. Such geographic location information may be used to describe the geographic location of the research subject of the article or the sampling point of the article's related research objects (as shown in Figure 7). The geographic location displayed on a map often does not display accurate latitude and longitude data. It is necessary to use the ruler or coordinate system of the map to read the relevant points to obtain the latitude and longitude values that can be stored. However, there are often many marked points on the map, and it is often difficult to convert the latitude and longitude coordinates of the pictures in the PDF. Researchers need to extract the map image from the PDF first and then use other tools, such as ArcGIS, to reconstruct the latitude and longitude coordinates of the map in order to read the accurate latitude and longitude data of the data points. Such a complex operational process currently requires the use of several different tools to perform, which is inefficient. Moreover, multiple conversions will also cause the accumulation of errors, which is not conducive to the accuracy of the final data.

2.3 | Summary

Due to the difference in encoding, version and source of PDF documents, the structure of internally stored digital information is chaotic and cannot be processed automatically by machines. Especially for some PDF files scanned from paper, the scanning quality dramatically affects the results of automated processing. Secondly, the scientific literature is a particular document presented in PDF format. Its purpose is not to present structured data but to express the author's research and opinions. Therefore, the data are scattered in different paragraphs and tables or figures according to different expression purposes, but not in a structured form. Based on the four different types of data we have discussed above and their distribution, we can understand the current challenges in extracting scientific data from the literature:

1. A large amount of articles need to be processed, but the manual extraction process faces a lot of mechanical labour (copy and paste as well as collation).
2. The complex structure of scientific literature makes searching for information difficult. The difficulty of extracting information is exacerbated by the different formats used by different journals.
3. The data in figures and tables cannot be extracted quickly due to the PDF format, and even with the help of tools, a complex process is required.

3 | OUR SOLUTION: GEODEEPSHOVEL

From February 2022, GeoDeepShovel has been deployed, and more than 40 geoscience research teams from geoscience departments of more than 10 universities are using GeoDeepShovel for data extraction and scientific database construction. As we mentioned in §2.1, to construct a scientific database, researchers need to process a large number of files and need to distribute files to team members for data extraction. It is a complex task for a research team to store and manage such a large number of documents. Considering that there are different members in the research team, the team needs to be divided into the data extraction task and the research team also needs to cooperate and communicate, team management is a considerable challenge.

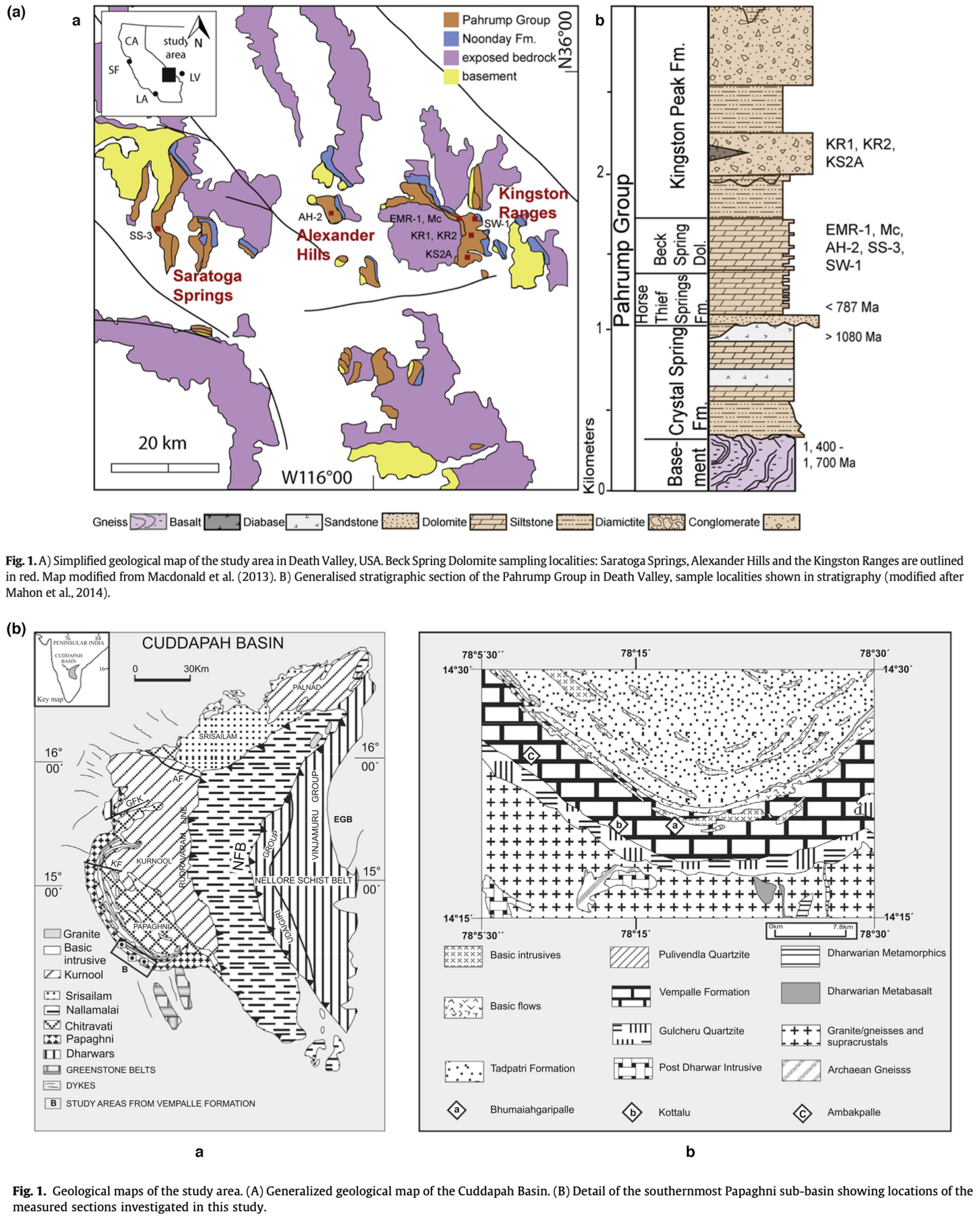


FIGURE 7 Different Map Style of Articles. (a) A map in The Tonian Beck Spring Dolomite: Marine dolomitization in a shallow, anoxic sea (Shuster et al., 2018). (b) A map in Carbonate platform development in a paleoproterozoic extensional basin, Vempalle formation, Cuddapah basin, India (Chakrabarti et al., 2014).

3.1 | GeoDeepShovel system overview

We solve these problems using a human–AI collaboration system design to extract data easily from PDF and have a better team cooperative experience. We design and implement some artificial intelligence modules for each task mentioned in §2.1. According to human–AI collaboration thinking, the system can collect the data for artificial intelligence model training while assisting humans in finishing the tasks. All data extracted by the user in the system are recorded as ground truth data, and the user's modification process is also recorded (e.g. modifying the value of a cell in a table), which will be used for fine-tuning and optimization of the model. Based on this process, the user's operation will not only complete the data extraction but also provide the relevant training data for the model. Such collaboration motivates people to participate and allows the machine to obtain enough information to improve.

As shown in Figure 8, GeoDeepShovel consists of: (1) an interactive graphical user interface (see Figure 9) including data extraction, document management, team management and data integration (D in Figure 9); (2) a back-end parse module to pre-process the PDF format files and (3) some back-end artificial intelligence models supporting data extraction and integration functions.

3.1.1 | Implementation details

The front-end interactive web application of GeoDeepShovel is developed in Vue.js and hosted with Nginx. The web-based design of GeoDeepShovel gives it the ability to run in web browsers on various platforms, including desktops, laptops, tablets and smartphones. The use of Vue.js and the design of a single-page application bring extreme load speed similar to native apps and consistent user experience across devices and platforms. The back-end API service of GeoDeepShovel is implemented with Python and FastAPI framework. The asynchronous coding design makes it possible to achieve higher concurrency with a minimal resource occupation so that it can support more users at the same time. We adopt a master–slave backup MySQL database to store documents and extracted data, which ensures data security and efficient reading and writing. Regarding user system security, we only store and bcrypt hashed passwords to ensure that users' plaintext passwords will not be stored and leaked. Moreover, the HTTPS protocol is applied to the whole system of GeoDeepShovel to ensure security in network communication.

3.2 | User system and document management

3.2.1 | Project and file management

In order to realize the management of projects, we designed the projects list interface to display the relevant information of each project, as shown in Figure 10a. Each project has a file list (see Figure 10b), which will show who uploaded the file, who was the last editor, the upload time and last edit time and whether the file has a principal. Users can change the project settings, including the text labels, the export dataset headers and the project description. Considering that the dataset may contain several headers, we provide batch editing for convenience. Moreover, to better browse and manage literature, the document list could be filtered with the principal user as well as the import user and sorted by title, import time and latest update time. To go further, each user could get "My File List" containing only documents taken charge of by him/her and "Recent File List" containing his recently viewed documents, which allow users to obtain the documents most important to them and simply continue their respective workflows.

Considering the particularity of system functions and the interaction process with the back-end model, we designed a file-locking mechanism to prevent more than one user from operating on the same file. Based on file lock, we implement a principal mechanism. Users can click the "Take Charge" button in the list to choose to be the "principal" in charge of a file. Then, the file can only be operated by the principal user, and other users can only read it. The user can release the file permission at any time.

3.2.2 | User system for team management

To help researchers manage their teams in data extraction work, we preset three team roles in the system design: Owner, Manager and Member. The user permissions of the three roles are shown in Table 1.

According to the users' description of their current team structure and cooperation, the users intend to ensure the original data's controllability and distinguish different research projects (the same team may carry out multiple projects). Therefore, in team management, we ban the modification of the project by the Member role to prevent the original data from being modified. At the same time, to ensure the rigour of the output dataset, we only open the modification of project settings to the Manager and

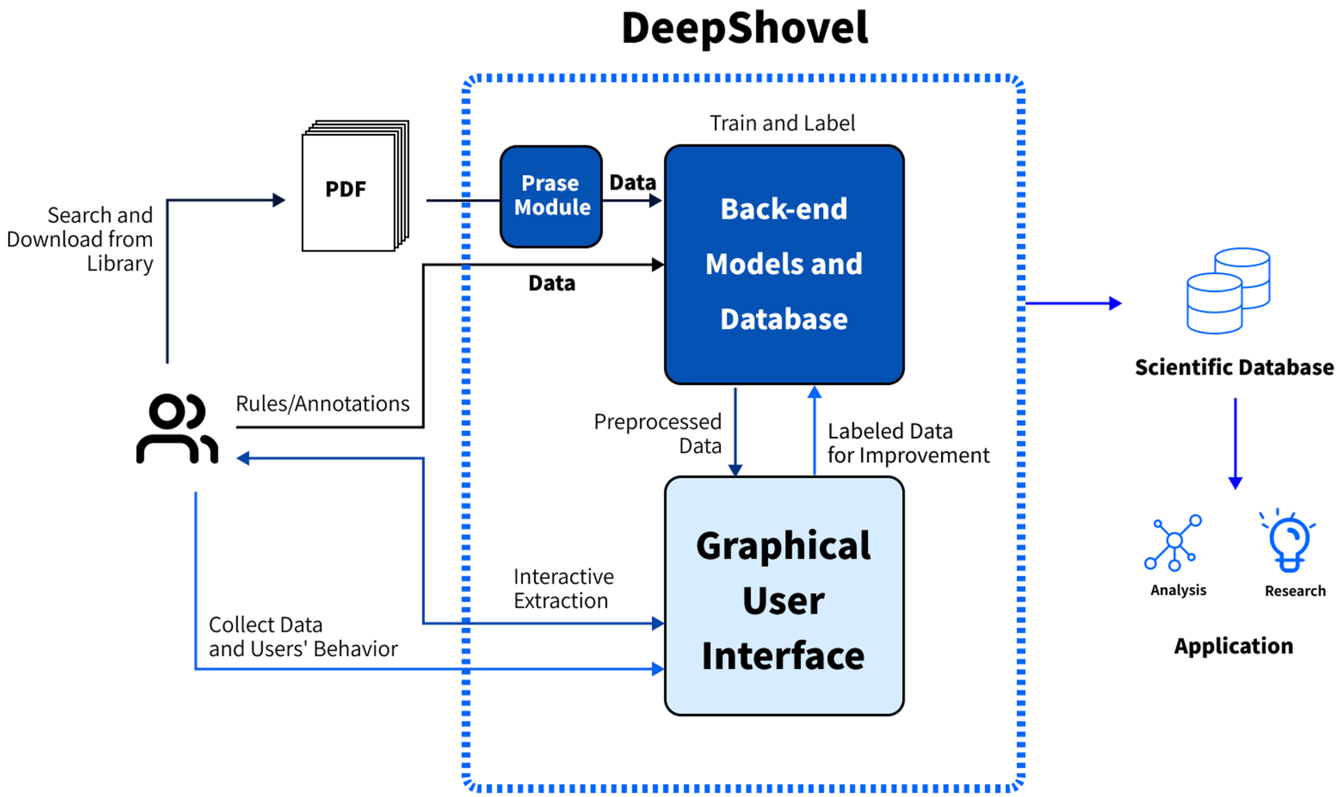


FIGURE 8 GeoDeepShovel System overview. GeoDeepShovel consists a PDF prasing module, a backend server including some artificial intelligence models and a interactive graphical user interface.

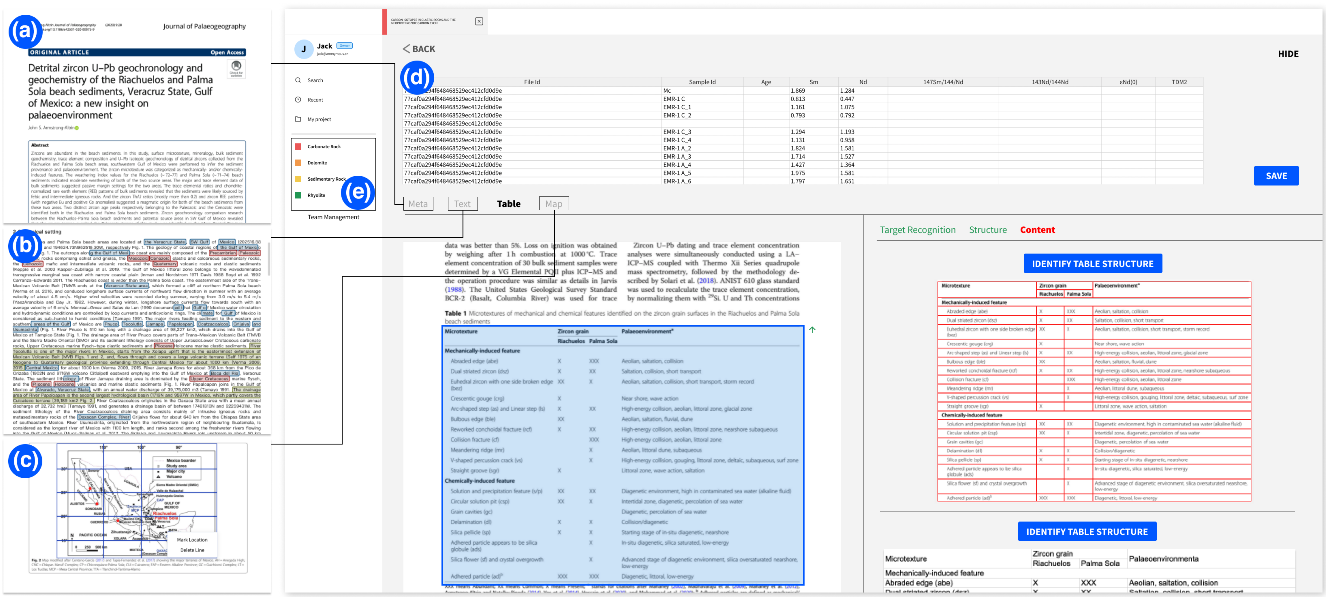


FIGURE 9 User Interface of GeoDeepShovel. The main part of this figure illustrates the table extraction and integration functions (d), while the system can also support meta information extraction (a), text extraction (b), map recognition and location extraction (c), and team and document management (e).

Owner. In order to meet the common scenario of cross-team collaboration in research, we allow users to join different teams. Team member removal does not affect his/her past actions.

3.3 | Data extraction process

All the literature uploaded into GeoDeepShovel is all automatically parsed with Grobid (GRO, 2008–2021) and

Science Parse (Tkaczyk et al., 2018). When users open a file from Project File List to start their work, they enter the data extraction interface (Figure 11). In the data extraction interface, users can switch the different tabs (e.g. Meta, Text, Table and Map) in the area F1. The details of each function are in the following sections.

3.3.1 | Metadata extraction

For each uploaded document, GeoDeepShovel uses multiple parsing tools (e.g. Grobid (GRO, 2008–2021), Science Parse and PdfFigures 2.0 (Clark & Divvala, 2016b)) to independently extract its meta-information and mix all the information with a voting mechanism. The meta-information of papers (e.g. Title, Author List, Abstract, Venue and Year) is extracted and indexed with Elasticsearch. Then, all the fields could be utilized for searching and retrieving the documents. As shown in Figure 11, users can edit and save the meta-information that can be joined to the output dataset.

3.3.2 | Name entity recognition and extraction from text

We use weak-supervision learning models and rules to help highlight the focused keywords in texts and the samples' features to help them add these words to the database. To extract academic entities from papers in the format of

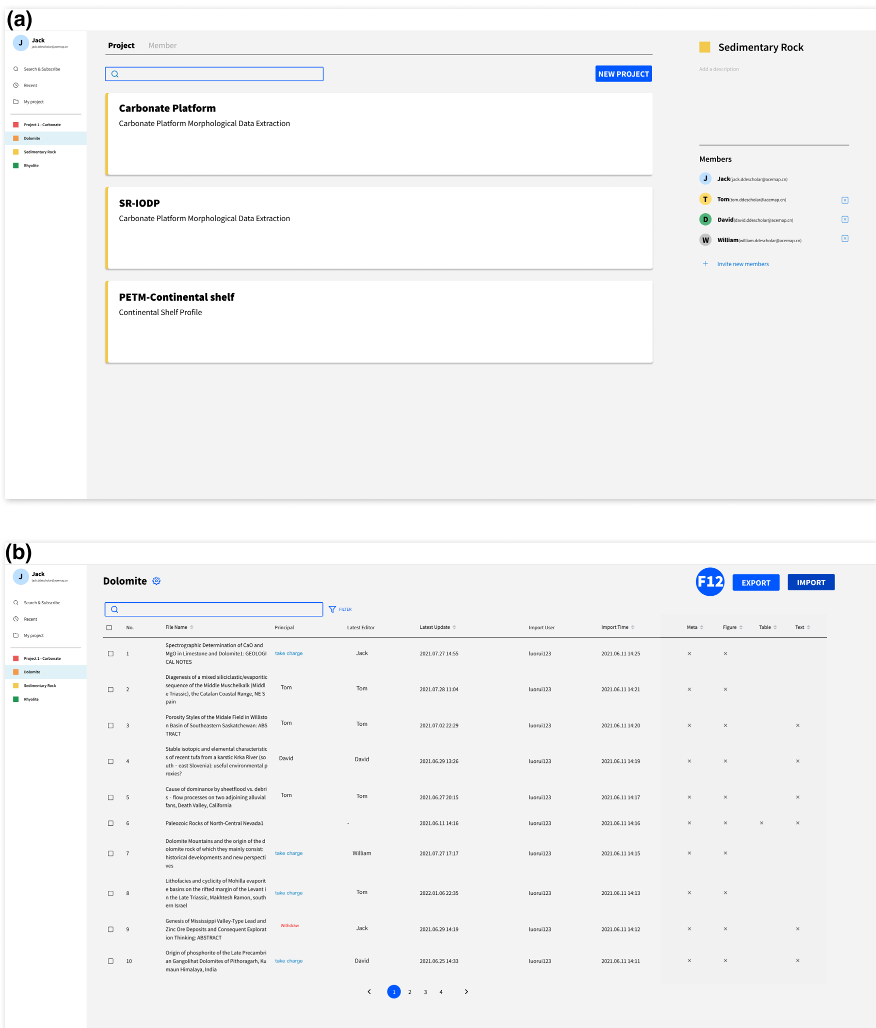


FIGURE 10 UI of Project and File Management. (a) Project List Page. User can see the project name and description. And user can delete the project here by right click. This page also consists the quickly team member add function. (b) File List Page. User can check the files' name and status here. And they can also use the "Take Charge" button to be a "principal" in charge of a file.

TABLE 1 The user permissions of each role.

Role	Add/remove manager	Add/remove member	Add/delete project	Import file	Project settings
Owner	√	√	√	√	√
Manager	-	√	√	√	√
Member	-	-	-	√	-

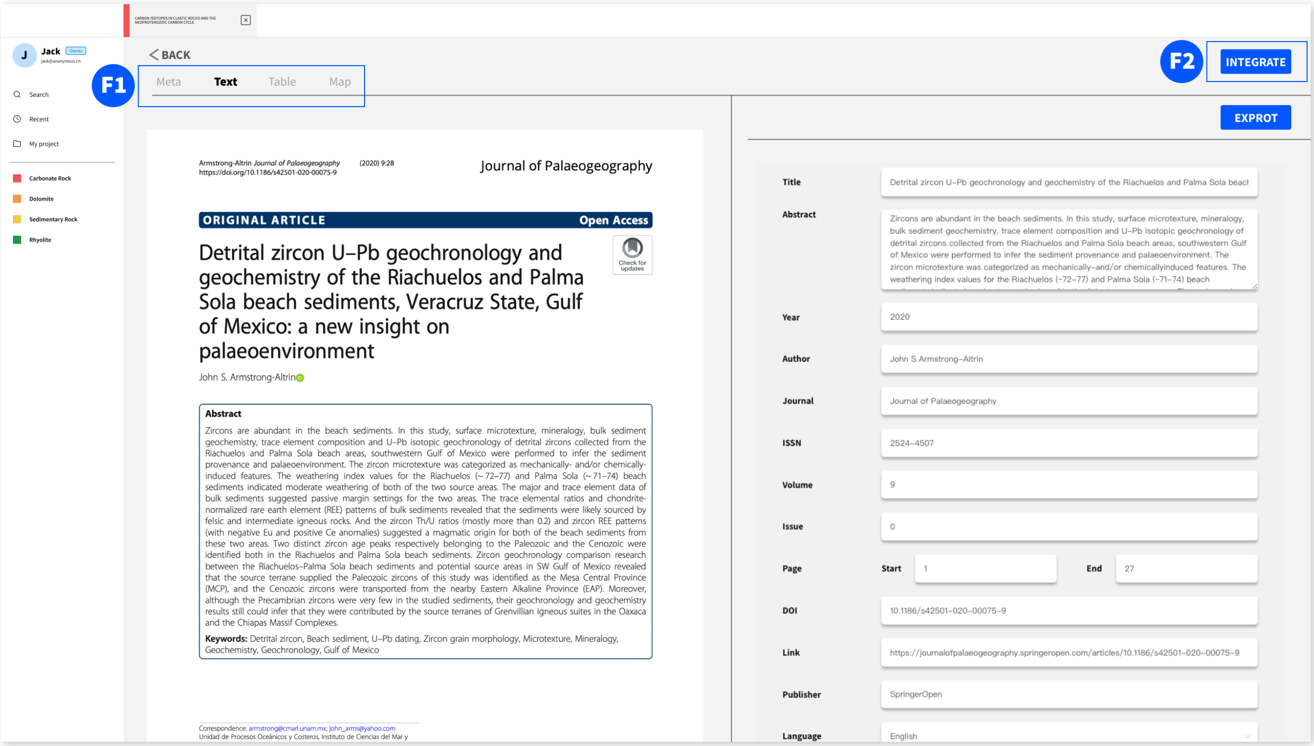


FIGURE 11 UI of Meta Information Extraction. User can edit the meta information here, which can also help the system to collect the correct information.

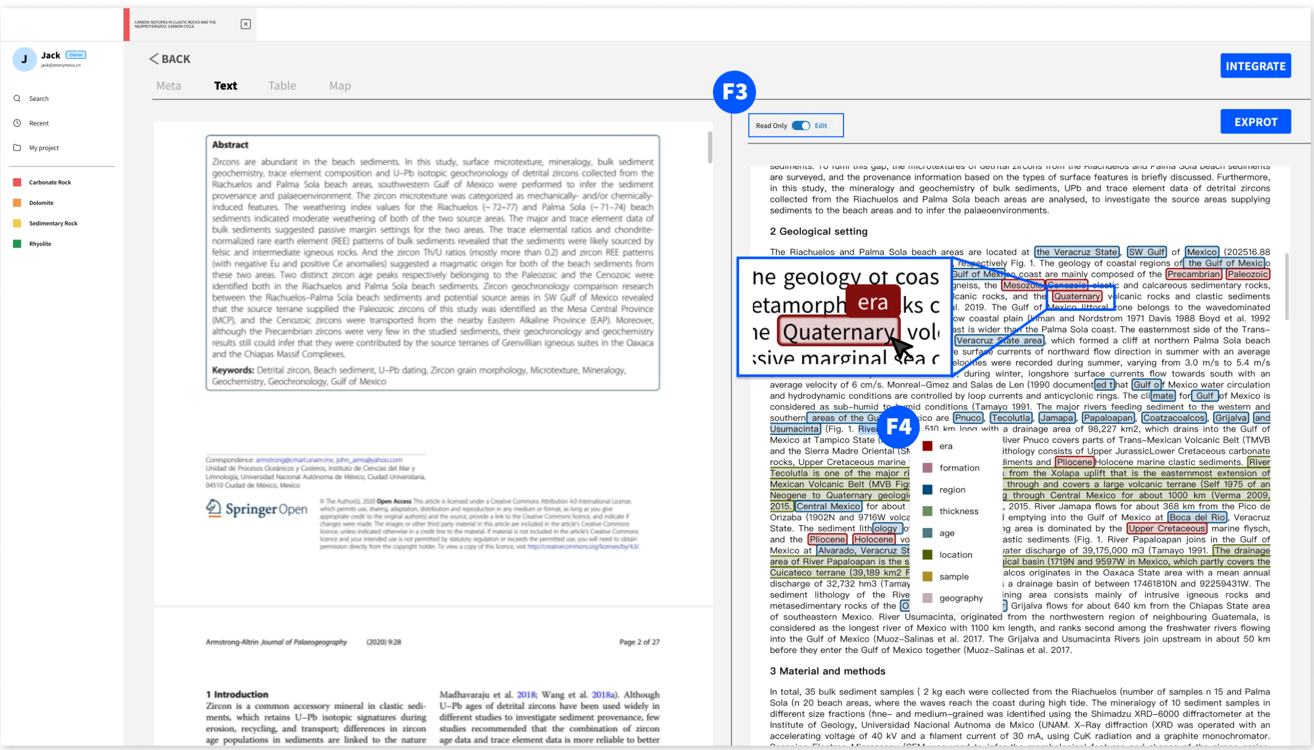


FIGURE 12 UI of Name Entity Recognition and Extraction. User can add/delete the entities and add them to the integrated table.

PDF, GeoDeepShovel first utilizes PDFFigures 2.0 (Clark & Divvala, 2016a) to parse each text section from the original files. Then, some rules and the natural language processing library spaCy (Honnibal & Montani, 2017) are adopted to automatically extract entities of different types from the parsed text sections.

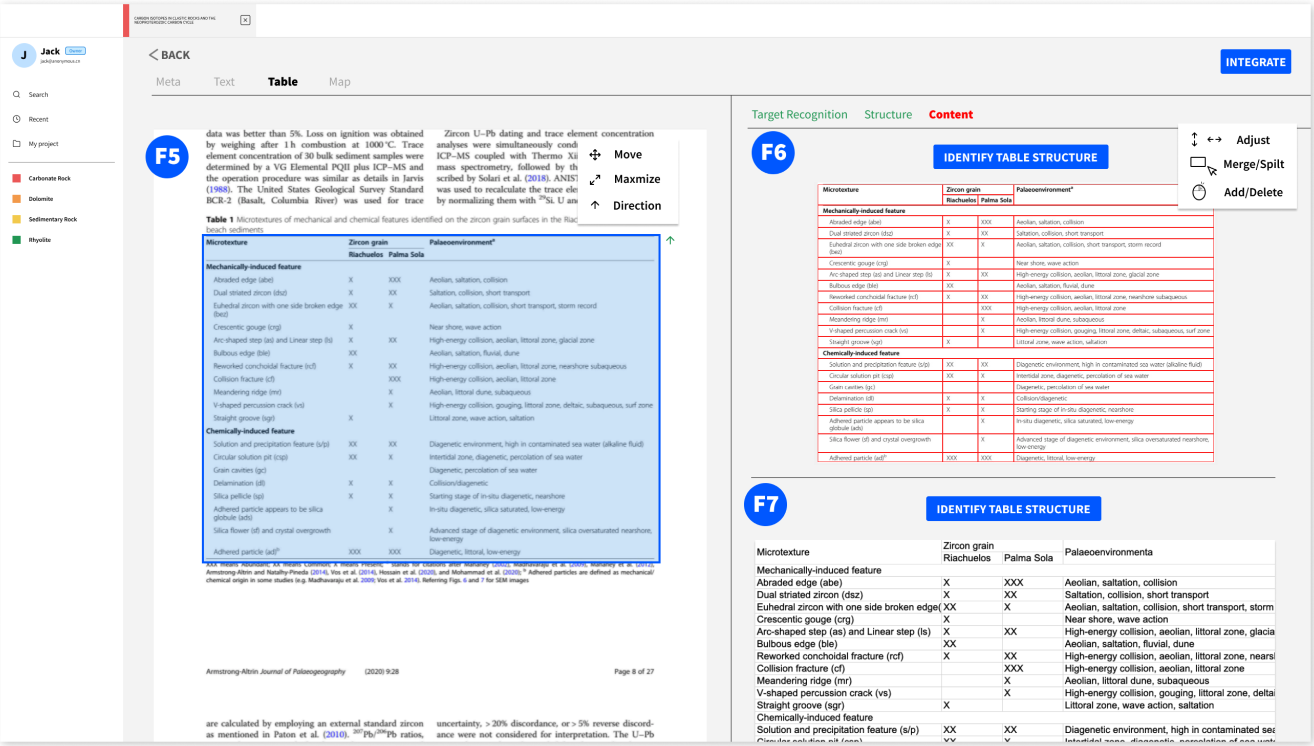


FIGURE 13 UI of Data Extraction from Table. User can adjust the structure and content of the table to get the correct data.

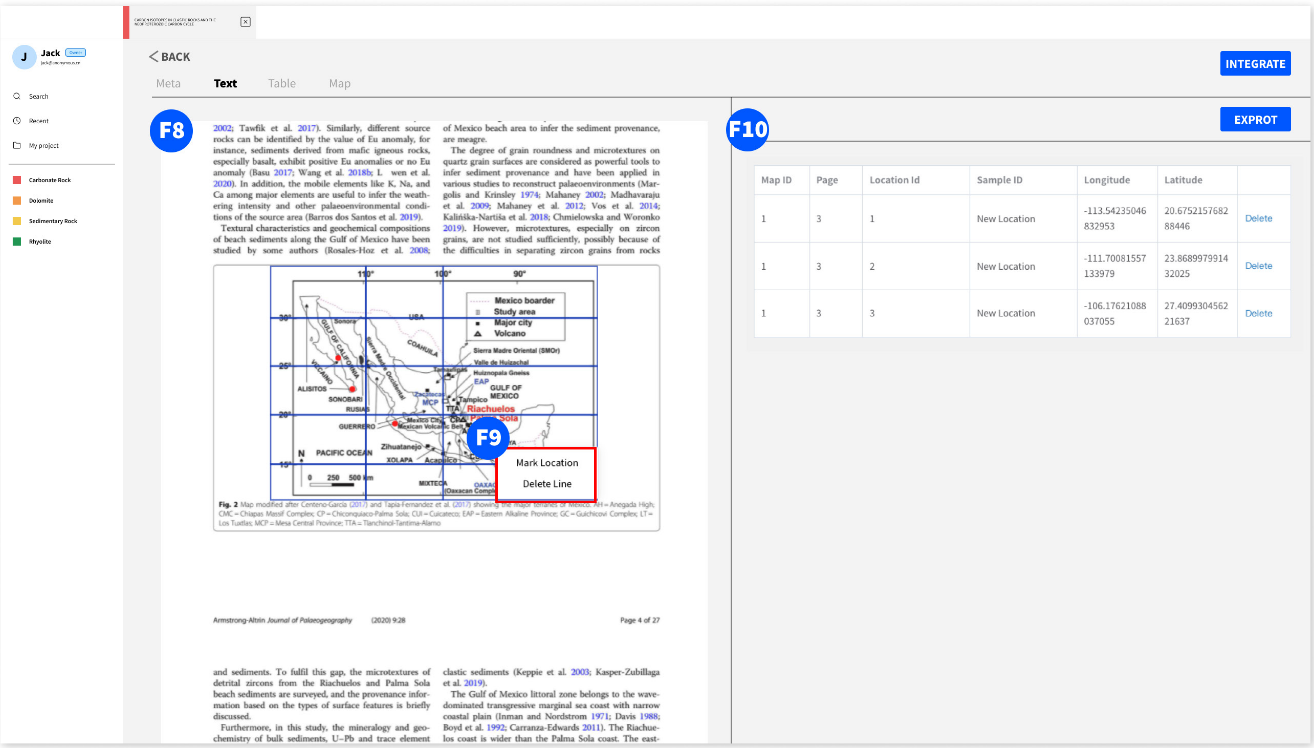


FIGURE 14 UI of Map Recognition and Location Extraction. User can adjust the recognized longitude and latitude line to get a correct coordinate range.

For example, we have a dictionary of eras' names to highlight the era mentioned in the PDF. Users can also annotate a keyword via mouse selection when they switch to the edit mode (F3) and select a label (F4) as shown in Figure 12. The keywords can be added to the output database (refer to Section 3.3.5). Users can choose to show or hide some labels, which are set at the project level as shown in Figure 15b.

3.3.3 | Data extraction from table

To help users extract the data in the table, we develop the Table Extraction function (Figure 13). First, GeoDeepShovel uses an object detection model Detectron2 (Wu et al., 2019) trained on TableBank (Li et al., 2019), a benchmark dataset for table detection, to detect the region of tables. Then for each table, a series of rules are adopted to locate each cell within it. Once users confirm the cell structure of a table, Tesseract (Kay, 2007) will be applied to detect the text in each cell and establish the final digitalized table.

In this part, we separate the task into three steps for the user to extract the table: (1) Locate the Table with the assistance of AI; (2) AI recognizes the table's structure, and users assist for a better result; (3) AI recognizes the table's content and users can edit for final accuracy. In each step, the artificial intelligence models we design will help people to easily get the result and collect the users' adjustments for model training. From the user's perspective, the first step is adjusting where the table is in the F5 area or drawing a new area as a table, then starting to recognize the structure. The next step is to adjust the structure that the system advised (F6). The system provides "add and delete column/row" and "merge or split cell function." After structure recognition, users can start the content recognition and edit the content in each cell (F7).

3.3.4 | Map recognition and location extraction

For collecting the location of a sample, we provide a module that can recognize maps and calculate the latitude and longitude of each point on the map (Figure 14). Users can draw an area (F8) that contains the map and mark a point by right click (F9). GeoDeepShovel will detect the longitude and latitude labelled at the map's margin and determine the map's coordinate range. Then, if users click any location on the map, the exact coordinates of the location will be automatically calculated and recorded. The latitude and longitude will automatically be saved in the table (F10) as shown in Figure 14 and can be joined to the output dataset (refer to Section 3.3.5).

3.3.5 | Data integration

After the data are extracted step by step, it needs to be integrated into a table to establish a database. We designed a single file integration and project-level integration with the assistance of AI to adapt to different teamwork modes. The user needs to set the header of the master table on the project page (Figure 15a). The header might be the same as the dataset's schema and contain the attributes extracted from different parts of the article. Then, in the data extraction interface (Figure 11), when users click the Integrate button (F2), the back-end model will process the data in each part, including meta-information, tables, location in maps and texts. In this process, the back-end server will automatically match the fields in the header of the master table with the headers of the sub-tables formed in the extraction of different parts, which will quickly bring all the data together. The result is shown in the F11 area in Figure 16.

After all the data in a single file have been integrated into a file-level summary table, the user can integrate the

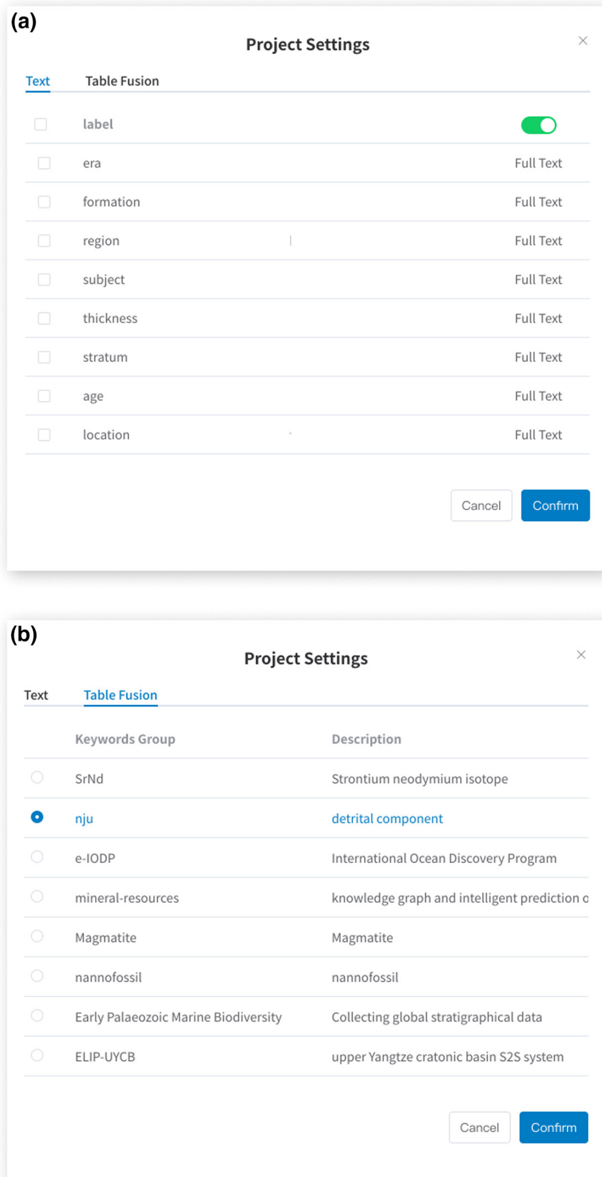


FIGURE 15 The Project Settings. (a) The settings of text extraction. (b) The settings of data integration.

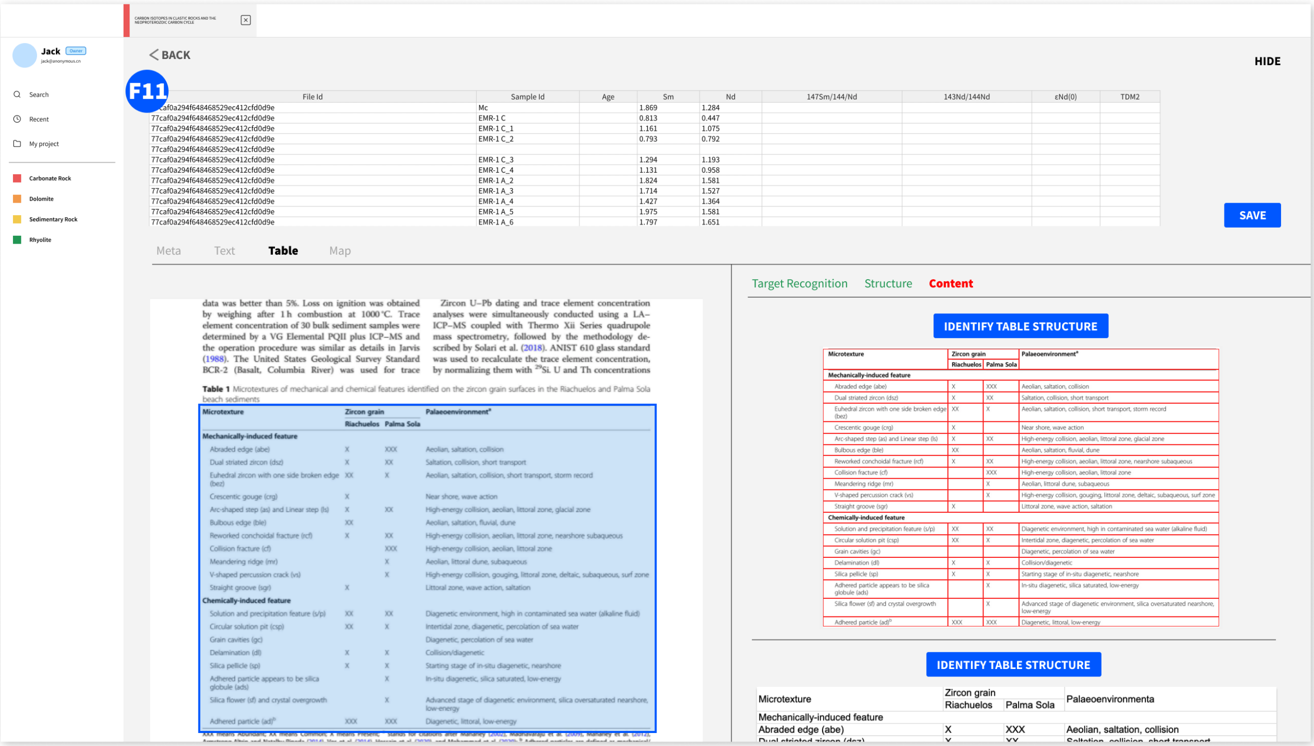


FIGURE 16 UI of Data Integration.

summary table of each file into a project-level summary table at the Files List interface (F12 in Figure 10b), and the result will be automatically downloaded.

3.4 | Example of data extraction workflow using GeoDeepShovel

Since building a database and conducting big data-driven scientific research is long-duration work, we show an example of a small dataset construction to demonstrate the capabilities of GeoDeepShovel. The dataset's schema is shown in Figure 17. Each data item is about the description of a bioevents. The researchers need to find out some attributes' values of the bioevents, including sample id (can be N/A), profile name, locality, latitude, longitude, age and depth.

We use the data extraction process from the article *Palynology of the Cenomanian to lowermost Campanian (Upper Cretaceous) Chalk of the Trunch Borehole (Norfolk, UK) and a new dinoflagellate cyst bioevent stratigraphy for NW Europe* (Pearce et al., 2020) to show how to build a database using GeoDeepShovel.

Usually, descriptions of the depth of bioevents and their ages are presented in a tabular form in such articles. Therefore, the first step in extracting data is to find tables containing biological events. GeoDeepShovel locates and highlights all tables in the article after the file has been uploaded to the system, allowing the user to quickly locate

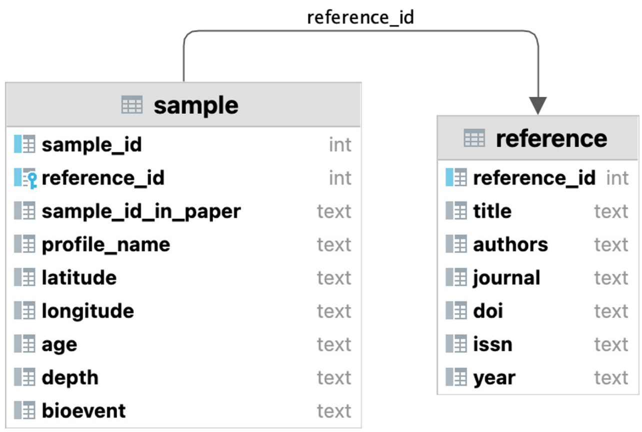


FIGURE 17 The UML of age model database. The age model database is used to build the age model for ocean drilling projects, including DSDP ODP and Two IODPs (at Scripps Institution of Oceanography, 2013-2022).

the table and confirm if it is the data to be extracted. Once it has been determined that the current table contains the required data, such as "Table 1 Age constraints for the Upper Cretaceous of the Trunch borehole," which contains details of the biological events, the user can click on the right-hand button to start identifying the structure in the table. After the user has made adjustments to the table structure, he/she can save and begin to identify the contents of the table and perform proofreading to ensure that the data are correct.

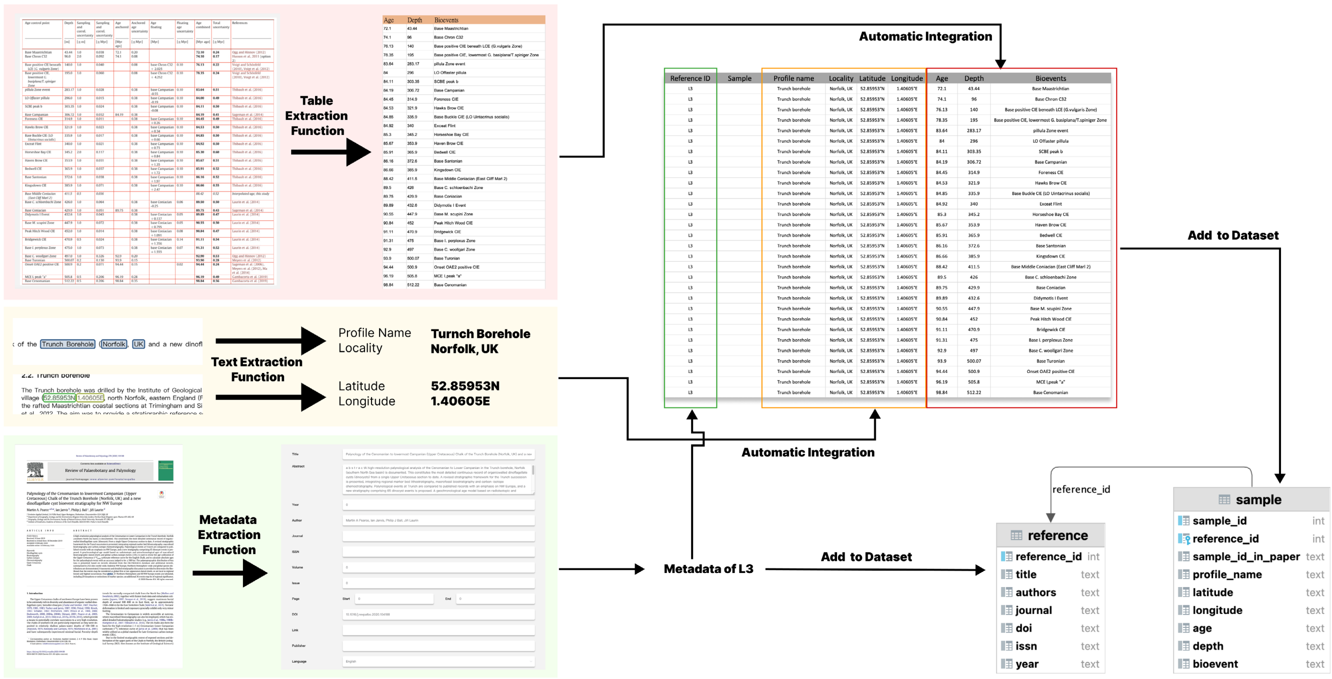


FIGURE 18 The example case workflow of using GeoDeepShovel.

The table contains the bioevents' names, depth and age in this example. Therefore, the rest of the data needs to be extracted from other parts of the article to complete the data. GeoDeepShovel provides a pre-label function in the text extraction process, which can identify and highlight the location name, latitude and longitude. This pre-label function is based on the parsing of PDF. Users can quickly target the desired entities or manually label and select an entity to add them to the final database. In this current task, the desired location name, latitude and longitude are already pre-marked and highlighted, so the user only needs to select to link.

After all types of data have been extracted, researchers can directly integrate all the data into a pre-set dataset structure. This step will significantly reduce the time spent in the workflow filling data into the forms one by one. Eventually, all data can be directly remitted to the dataset. The complete process is shown in Figure 18.

4 | CONCLUSION AND FUTURE WORKS

In this paper, we present GeoDeepShovel, an online platform for data extraction in the scientific literature with AI assistance that can help researchers cooperate with their teammates to extract data from PDF documents and build a scientific database. GeoDeepShovel can help researchers extract and aggregate data containing meta-information, tables, texts and location from the literature. The research team can collaborate in GeoDeepShovel, and team members can share resources and progress with others.

We propose a novel and general collaborative framework for scientific literature data extraction in geoscience, which can help with big data-driven discovery. By extracting fine-grained data from text, tables and images in PDFs, the constructed database can cover different dimensions more completely, which is beneficial to the subsequent data analysis and modelling process. In GeoDeepShovel, the researcher makes the final decision of all data extraction, and AI fully follows the user's instructions in this interaction process to ensure the accuracy of the data. The reason we designed this workflow is that due to the accumulation of errors in end-to-end approaches, the final outcome might be unacceptable for scientific research. Meanwhile, manually checking and correcting these errors is a very tedious and difficult job. We would like to claim that we do not believe today's AI technology can build a “fully automated” system to replace researchers in data extraction. To ensure the quality of the database used in further research, researchers still have to clean and correct data manually. We think that building a human–AI collaboration solution with the appropriate level of automation would be a better way to solve the problem so that the human and AI can jointly iterate, improve and complete the data extraction.

There are still some technical limitations in GeoDeepShovel: (1) the quality of data extraction is greatly affected by the quality of PDF files, and we cannot handle some low-resolution scans that are too old; (2) most AI models in GeoDeepShovel are based on rules provided by geoscientists and a relatively small amount of geoscience data, which may lead to some problems in the processing of uncovered literature; (3) in the current proof-of-concept stage, GeoDeepShovel does not meet all the types of data demands (e.g. points location in some scatterplots) because of the lack of relevant ground truths.

In the future, we will continuously add new modules and improve existing modules through rapid system iterative upgrades.

AUTHOR CONTRIBUTIONS

Shao Zhang: Conceptualization (equal); data curation (equal); investigation (equal); methodology (equal); validation (lead); visualization (lead); writing – original draft (lead). **Hui Xu:** Conceptualization (equal); software (equal); writing – review and editing (equal). **Yuting Jia:** Conceptualization (equal); data curation (equal); investigation (equal); software (lead); writing – original draft (equal). **Ying Wen:** Conceptualization (equal); formal analysis (equal); methodology (equal); project administration (equal); supervision (lead); writing – review and editing (equal). **Dakuo Wang:** Investigation (equal); methodology (equal); supervision (equal); writing – review and editing (equal). **Luoyi Fu:** Supervision (equal). **Xinbing Wang:** Funding acquisition (equal); project administration (equal); supervision (equal). **Chenghu Zhou:** Funding acquisition (equal).

ACKNOWLEDGEMENTS

We owe a particular debt of gratitude to the scientists from the Deep-time Digital Earth program who all contributed enormously valuable feedback. We also thank Jia Guo, Yifei Shen, Qi Li, Zhixin Guo, Mingxuan Yan, Mingze Li, Le Zhou, Jingyao Tang, Han Liu, Shengling Zhu and Tao Shi from IIOT Research Center in Shanghai Jiao Tong University for their support to our system development. This work is supported by the National Natural Science Foundation of China (No.42050105, No.62106141) and the Shanghai Sailing Program (21YF1421900). This work is a part of the Deep-time Digital Earth (DDE) Big Science Program.

FUNDING INFORMATION

This work is supported by the National Natural Science Foundation of China, Grant Number: No.42050105 and No.62106141 and Shanghai Sailing Program, Grant Number: 21YF1421900.

OPEN RESEARCH BADGES

This article has earned Open Data, Open Materials and Preregistered Research Design badges. Data, materials and the preregistered design and analysis plan are available at Open Science Framework

ORCID

Shao Zhang https://orcid.org/0000-0002-0111-0776
Ying Wen https://orcid.org/0000-0003-1247-2382

REFERENCES

Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P. et al. (2019) *Guidelines for human-AI interaction, page 1–13*. New York, NY, USA: Association for Computing Machinery. Available from: https://doi.org/10.1145/3290605.3300233

Armstrong-Altrin, J.S. (2020) Detrital zircon u–pb geochronology and geochemistry of the riachuelos and Palma Sola beach sediments, Veracruz state, gulf of Mexico: A new insight on palaeoenvironment. *Journal of Palaeogeography*, 9(1), 1–27.

Ashktorab, Z., Desmond, M., Andres, J., Muller, M., Joshi, N.N., Brachman, M. et al. (2021) Ai-assisted human labeling: Batching for efficiency without overreliance. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1), 1–27. Available from: https://doi.org/10.1145/3449163

Bergen, K.J., Johnson, P.A., Maarten, V. & Beroza, G.C. (2019) Machine learning for data-driven discovery in solid earth geoscience. *Science*, 363(6433), eabg9551.

Brand, L., Wang, M. & Chadwick, A. (2015) Global database of paleocurrent trends through the phanerozoic and precambrian. *Scientific Data*, 2(1), 1–7.

Cervato, C., Bohling, G., Loepp, C., Taylor, T., Snyder, W.S., Diver, P., Reed, J., Fils, D., Greer, D., and Tang, X. (2005) The chronos system: Geoinformatics for sedimentary geology and paleobiology. In 2005 IEEE international symposium on mass storage systems and technology, pp. 182–186. IEEE.

Chakrabarti, G., Shome, D. & Kumar, S. (2014) George M Stephens III, and Linda C Kah. Carbonate platform development in a paleoproterozoic extensional basin, vempalle formation, Cuddapah basin, India. *Journal of Asian Earth Sciences*, 91, 263–279.

Clark, C. and Divvala, S. (2016a) Pdffigures 2.0: Mining figures from research papers. In 2016 IEEE/ACM joint conference on digital libraries (JCDL), pp. 143–152. IEEE.

Clark, C. & Divvala, S. (2016b) Pdffigures 2.0: Mining figures from research papers.

Desmond, M., Muller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K. et al. (2021) *Increasing the speed and accuracy of data labeling through an AI assisted Interface*. New York, NY, USA: Association for Computing Machinery, pp. 392–401. Available from: https://doi.org/10.1145/3397481.3450698

Dirzo, R., Young, H.S., Galetti, M., Ceballos, G., Isaac, N.J.B. & Collen, B. (2014) Defaunation in the anthropocene. *Science*, 345(6195), 401–406.

Fan, J.-X., Shen, S.-Z., Erwin, D.H., Sadler, P.M., MacLeod, N., Cheng, Q.-M. et al. (2020) A high-resolution summary of cambrian to early triassic marine invertebrate biodiversity. *Science*, 367(6475), 272–277.

Govindaraju, V., Zhang, C., and Ré, C. Understanding tables in context using standard nlp toolkits. In Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 2: Short papers), pp. 658–664, 2013.

Grobid. (2008–2021) https://github.com/kermitt2/grobid.

Hoeppe, G. (2021) Encoding collective knowledge, instructing data reusers: The collaborative fixation of a digital scientific data set. *Computer Supported Cooperative Work (CSCW)*, 30(4), 463–505.

Honnibal, M. & Montani, I. (2017) spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Kay, A. (2007) Tesseract: An open-source optical character recognition engine. *Linux Journal*, 2007(159), 2.

Li, M., Cui, L., Huang, S., Wei, F., Zhou, M. & Li, Z. (2019) Tablebank: A benchmark dataset for table detection and recognition.

McDowell, R.W., Noble, A., Pletnyakov, P. & Mosley, L.M. (2021) Global database of diffuse riverine nitrogen and phosphorus loads and yields. *Geoscience Data Journal*, 8(2), 132–143.

McMahon, W.J. & Davies, N.S. (2018) Evolution of alluvial mudrock forced by early land plants. *Science*, 359(6379), 1022–1024.

National Research Council, Division on Engineering and Physical Sciences, Commission on Physical Sciences, Mathematics, and Applications, Committee for a Study on Promoting Access to Scientific and Technical Data for the Public Interest. (2000) *A question of balance: Private rights and the public interest in scientific and technical databases*. Washington, DC: National Academies Press.

Niu, F., Zhang, C., Ré, C. & Shavlik, J.W. (2012) Deepdive: Web-scale knowledge-base construction using statistical learning and inference. *VLDS*, 12, 25–28.

Oberhänsli, R. (2020) Deep-time digital earth (dde) the first iugs big science program. *Journal of the Geological Society of India*, 95(3), 223–226.

Parés, J.M., Vernet, E., Calvo-Rathert, M., Soler, V., Bógalo, M.-F. & Álvaro, A. (2022) Rock magnetism of lapilli and lava flows from cumbre vieja volcano, 2021 eruption (la Palma, canary islands): Initial reports. *Geosciences*, 12(7), 271.

Pearce, M.A., Jarvis, I., Ball, P.J. & Laurin, J. (2020) Palynology of the cenomanian to lowermost campanian (upper cretaceous) chalk of the trunch borehole (Norfolk, UK) and a new dinoflagellate cyst bioevent stratigraphy for nw europe. *Review of Palaeobotany and Palynology*, 278, 104188.

Puetz, S.J. (2018) A relational database of global u–pb ages. *Geoscience Frontiers*, 9(3), 877–891. Available from: https://doi.org/10.1016/j.gsf.2017.12.004 https://www.sciencedirect.com/science/article/pii/S1674987117302141. Greenstone belts and their mineral endowment

Puetz, S.J., Ganade, C.E., Zimmermann, U. & Borchardt, G. (2018) Statistical analyses of global u-pb database 2017. *Geoscience Frontiers*, 9(1), 121–145.

Renaudie, J., Lazarus, D.B. & Diver, P. (2020) Nsb (neptune sandbox berlin): An expanded and improved database of marine planktonic microfossil data and deep-sea stratigraphy. *Palaeontologia Electronica*, 23, a11.

Science Support Office at Scripps Institution of Oceanography. (2013-2022) The international ocean discovery program (iodp). https://www.iodp.org/

Shuster, A.M., Wallace, M.W., van Smeerdijk Hood, A. & Jiang, G. (2018) The tonian beck spring dolomite: Marine dolomitization in a shallow, anoxic sea. *Sedimentary Geology*, 368, 83–104.

Snyder, W.S., Lehnert, K.A., Ito, E., Harms, U., and Klump, J. (2008) Geoscinet: Building a global geoinformatics partnership. In AGU fall meeting abstracts, vol 2008, pp. IN31D–03.

Sun, Z., Sandoval, L., Crystal-Ornelas, R., Mousavi, S.M., Wang, J., Lin, C. et al. (2022) A review of earth artificial intelligence. *Computers & Geosciences*, 159, 105034.

Tkaczyk, D., Collins, A., Sheridan, P. & Beel, J. (2018) Machine learning vs. rules and out-of-the-box vs. retrained: An evaluation of open-source bibliographic reference and citation parsers.

Tucker, M.A., Böhning-Gaese, K., Fagan, W.F., Fryxell, J.M., Van Moorter, B., Alberts, S.C. et al. (2018) Moving in the anthropocene: Global reductions in terrestrial mammalian movements. *Science*, 359(6374), 466–469.

Wang, C., Hazen, R.M., Cheng, Q., Stephenson, M.H., Zhou, C., Fox, P. et al. (2021) The deep-time digital earth program: Data-driven discovery in geosciences. *National Science Review*, 8(9), nwab027. Available from: https://doi.org/10.1093/nsr/nwab027

Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A. et al. (2016) The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), 1–9.

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y. & Girshick, R. (2019) Detectron2. https://github.com/facebookresearch/detectron2

Zhang, C., Govindaraju, V., Borchardt, J., Foltz, T., Ré, C. & Peters, S. (2013) Geodeepdive: Statistical inference using familiar data-processing languages. In: *Proceedings of the 2013 ACM SIGMOD international conference on Management of Data, SIGMOD '13*. New York, NY, USA: Association for Computing Machinery, pp. 993–996. Available from: https://doi.org/10.1145/2463676.2463680

How to cite this article: Zhang, S., Xu, H., Jia, Y., Wen, Y., Wang, D., Fu, L. et al. (2023) GeoDeepShovel: A platform for building scientific database from geoscience literature with AI assistance. *Geoscience Data Journal*, 00, 1–19. Available from: https://doi.org/10.1002/gdj3.186