DOI: 10.1002/gdj3.186

DATA SERVICES ARTICLE



WILEY

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

Shao Zhang¹ | Hui Xu¹ | Yuting Jia¹ | Ying Wen¹ | Dakuo Wang² | Luoyi Fu¹ | Xinbing Wang¹ | Chenghu Zhou³

¹Shanghai Jiao Tong University, Shanghai, China ²IBM Research, Cambridge, Massachusetts, USA

³Institute 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.

K E Y W O R D S

artificial intelligence, big data-driven discovery, data extraction, scientific database, humancomputer 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 datadriven 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

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2023 The Authors. Geoscience Data Journal published by Royal Meteorological Society and John Wiley & Sons Ltd.

RMetS

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 endto-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

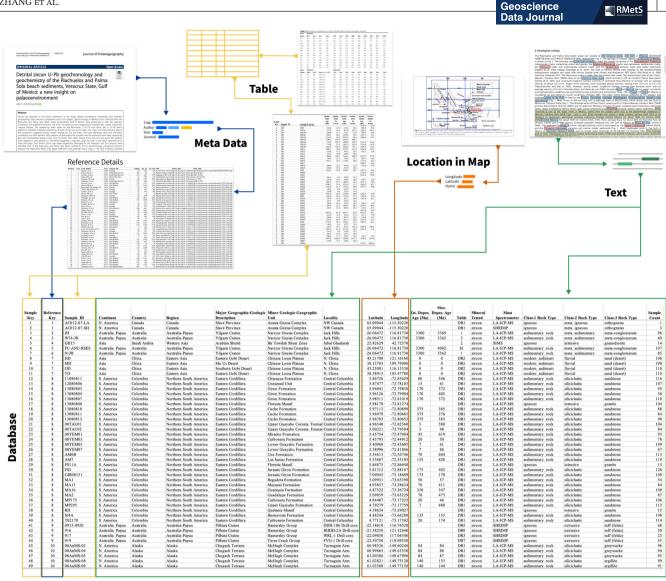


FIGURE 1 A workflow's example for extracting data from a literature and save to database. The researchers extract the corresponding values of the attributes of the samples from different parts of the article and fill them in the dataset. (Figures and table are from (Armstrong-Altrin, 2020). The database 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

3

(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.

RMetS 🖉

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, wordoriented 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' metainformation 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

Iters

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).

Geoscience

Data Journal

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 metainformation 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

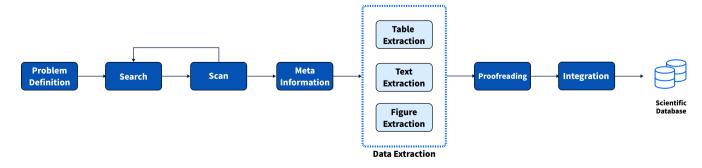


FIGURE 3 The team's workflow of collecting and building a database from geoscience literature.

5

RMetS

6

RMetS

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

(a)

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

Journal of Palaeogeography

Access

ORIGINAL ARTICLE

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 detrial aircons collected from the Rischuelos and Palma Sola beach areas, southwestern Gulf of Mexico were performed to infer the sediment provenance and palaecenvironment. The zircon microtexture was categorized as mechanically- and/or chemically induced features. The weathering index values for the Rischuelos (~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 extinement. The strategies may another the for the hole one. The trace indement police and the observed in the sediments indicated moderate weathering of both of the two source areas. The major and trace element data of the hole of the two sources. 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 aircon Th/J ratios (mostly more than 0.2) and zircon REE patterns in the weather 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-Palma Sola beach sediments. Zircon gecknology 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. Cental 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 and the Chipapa Massif Complexes. **Rewords:** Detrification provides the studied sediments for the maters of Grenvillian igneous suites in the Oaxac and the Chipapa Massif Complexes.

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

Deringer Open

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.

(b) RESEARCH

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

Jun-suan Fan¹⁴, Shu-zhong Shen^{1,23}, Douglas H. Erwin^{4,5} Peter M. Sateler⁴, Norman M. Qiu-ming Cheng², Xu-dong Hou², Jiao Yang², Xiang-dong Wang², Yue Wang², Hua Zhang², Gao-Jiang Ji², Yi-dhan Zhang², Yu-kun Shi⁴, Dong-cun Yuan², Qing Chen², Lin-na Zhang² Chou L², Ying² Jing² Zhou²

One great challenge in understanding the history of life is resolving the influe One great challenge in understanding the history of life is releving the influence of environmental change on biodiversity. Simulated amounting and genetic algorithms were used to synthesize data hard 11.000 minime fossil species, collected from more than 3000 stratigraphic sections. Is generate a new Cambrian to Triasso biodiversity curve with an impade temporal resolution of 28 H3 thosand years. This increased resolution califies the timing of known diversity and increases and the strategies analysis suggests that partial pressure of carlon disolog (Pos) bit end y environmental factor that second to doubly a sinuality has the bias of bias in the site of the site

nderstanding patterns of global diversity can reveal the history of the biosphere and relations between environmental damages and diversity fluctuations, and can provide insights into how the fossil record might inform current biodiversity con-erns. Early global-scale quantitative analysis dentified what have come to be known as the identified what have come to be known as the "big five" mass extinctions. However, such ef-forts depend on the quality and temporal resolution of paleontological data, which have improved substantially since the 1990s, most recently through the intensive data compilation of the Paleobiology Database (1-6). Analyses

of the Paleobiology Database (1–6). Analyses of those data have increased our understand-ing of paleobiodiversity (δ , 7–10). Previous deep-time paleobiodiversity recon-structions (I, II) were limited by coarse age determinations of taxon occurrences. The relatively long and uneven duration of age bins (stage or series level) used in these studies imposed complexly structured limits on resolv ing power across different intervals. Reso generally no better than 8 to s (Myr) with standard deviations of 2.4 to 3.2 Myr, although some trials have been made to achieve better resolution for the

th Sciences and Engineering, Nanjing njing 210023, China. ²LPS, Nanjing Ir

Fan et al., Science 367, 272-277 (2020) 17 January 2020

early Palecoxic (d). Taxon age assignments were subject to error, not equally applicable to all clades, and quickly became outlated by new correlations or updated age estimates. Pro-vious analyses have also been performed at taxonomically broad and phylogenetically auspect family or genus levels. Such resolution are often to or rule and imprecise to assess diversification rates or patterns associated with various global events (gradual, stepwise, or abrupt) and may mask multiple events as well as finer-scale fluctuations (7, 12, 13). Here, we used a new parallel computing im-plementation of the constrained optimization

method (CONOP.SAGA) run on the Tianhe II supercomputer. This approach uses inferred stratigraphic correlations to construct compo-site biodiversity curves for Cambrian to Triassic marine invertentate genera and species (Fig. 1) and has demonstrated the capacity to establish finely resolved, traceable time zones over wide geographic areas (14).

Data and methods

Data compilation and standardization were conducted through the Geobiodiversity Data (15). This database is particularly suitable for biodiversity studies because, unlike the Paleo-biology Database (Fig. 2), it is based on section data and provides quality control at a bed-by-bed level using an online, interactive system bed level using an online, interactive system for recording expert taxonomic opinions (75). Taxonomic and age assignments used in this investigation were vetted by a team of 11 paleontologists, who checked and updated ach toxonomic record. We also cross-scheded the spectra in the for symmyons. Bornis is a strateging of the strateging of the sitions in stratigraphic sections, we were able to construct a composite sequence of assemb-

lages and callbrate this sequence to a current lages and callbrate this sequence to a current best available chronostruitgraphic data (*D*). Our study focused on marine invertebrates and used data from 3760 published strati-gords of the stratingraphic ranges of 46.318 toxonomic units, covering all Chinese Combrino to Lower Thisse tectorine blocks on which they reside were situated in paleolatitudes stretching from southern Gondarian to north-em Borneal realms (*T*). Accordingly, these data reflect global biodiversity patterns (figs. 52) ern Boreal realms (17). Accordingly, these data reflect global biodiversity patterns (figs. S2

reflect global inouverses parameters and S3). Our initial analyses revealed that the rarity of Silurian-Devonian data in China (due to wordwide regression) hampered regional and global correlations. Consequently, we added a small amount of European Silurian-Devonian data to improve the correlations in this in-terval. These additional data this hot alter the generality of our results because different buckets were blockets in difficult in different to the second in the second in different data of the second buckets were blockets were blocket in different to the second in the second in different data of the second in different to the second in the second in different data of the second in different to the second data of the second data of the second data of the second term of the second data of the se the generality of our results because different Chinese tectonic blocks were located in dif-ferent regions during the Paleozoic, with some residing close to Europe (fig. S3). Our study interval terminated at the late Middle Triassic marine nergenier.

narine regression. Taxonomic names in open nomenclature, questionable taxa, and taxa unidentifiable to questionable task and task informations of the species level were not included. Species recorded from only one locality were also removed to avoid the "monograph effect" (18). The resulting final dataset contained 116,060 local records of total stratigraphic ranges of 11,268 species from 3112 published strati

11,268 species from 312 published strati-graphic sections. To avoid the need to use coarse time bins, we used constrained optimization (CONOP) (19) stratigraphic correlation to reconstruct the Paleozoic biodiversity history of marine invertebrates. The CONOP correlation meth-al ability of the participation of the participation of the pathology of the participation of the pathology of the p invertebrates. The CONOP correlation mean od, which applies a simulated annealing al-gorithm to infer a globally optimized sequence of stratigraphic datums, has been used previ-ously for local high-resolution biochronostrat-tic strate (14: 90 - 90). However, the igraphic studies (14, 20, 21). However, the original CONOP algorithm (22, 23) did not support parallel or high-performance com-puting and it would have required dozens of years to calculate one CONOP composite for this dataset. To overcome this "big data" problem, we modified the original CONOP algorithms to parallelize the original Corporation of the sequencing prob-lem. We also designed a special hybrid strategy of simulated annealing and genetic algorithm for the parallel computing application, CONOP. SACA (15) CONOP.SAGA iteratively compares species

anges from many local range charts to as-semble the global first and last occurrence datums into a single, global, best-fit sequence, thereby reducing the effect of local-section

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 grai	n	Palaeoenvironment ^a
	Riachuelos	Palma Sola	
Mechanically-induced feature			
Abraded edge (abe)	Х	XXX	Aeolian, saltation, collision
Dual striated zircon (dsz)	Х	XX	Saltation, collision, short transport
Euhedral zircon with one side broken edge (bez)	XX	Х	Aeolian, saltation, collision, short transport, storm record
Crescentic gouge (crg)	Х		Near shore, wave action
Arc-shaped step (as) and Linear step (ls)	Х	XX	High-energy collision, aeolian, littoral zone, glacial zone
Bulbous edge (ble)	XX		Aeolian, saltation, fluvial, dune
Reworked conchoidal fracture (rcf)	Х	XX	High-energy collision, aeolian, littoral zone, nearshore subaqueous
Collision fracture (cf)		XXX	High-energy collision, aeolian, littoral zone
Meandering ridge (mr)		Х	Aeolian, littoral dune, subaqueous
V-shaped percussion crack (vs)		Х	High-energy collision, gouging, littoral zone, deltaic, subaqueous, surf zone
Straight groove (sgr)	Х		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	Х	Intertidal zone, diagenetic, percolation of sea water
Grain cavities (gc)			Diagenetic, percolation of sea water
Delamination (dl)	Х	Х	Collision/diagenetic
Silica pellicle (sp)	Х	Х	Starting stage of in-situ diagenetic, nearshore
Adhered particle appears to be silica globule (ads)		Х	In-situ diagenetic, silica saturated, low-energy
Silica flower (sf) and crystal overgrowth		Х	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
	TP	2,268 (80,120)	640 (40,897)	254	Programme (2018)
	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);
	TP	723 (74,571)	449 (31,121)	311	Larned et al. (2016)
	NO _x -N	723 (74,573)	449 (31,121)	86	
	TN	723 (74,573)	449 (31,121)	26	

^aOnly used as part of the technical validation.

RMetS

Geoscience

Data Journal



RMetS

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.

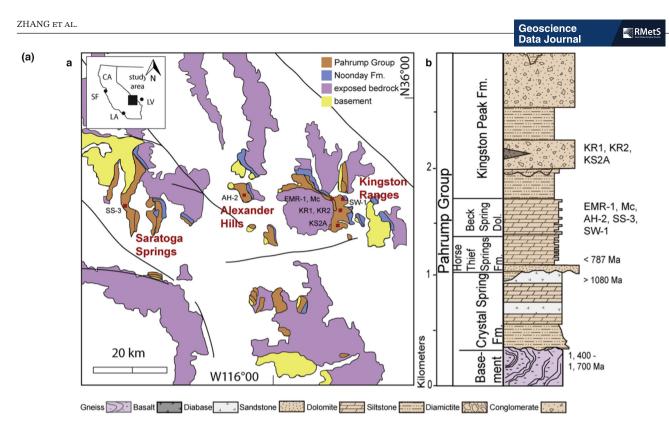


Fig. 1. A) Simplified geological map of the study area in Death Valley, USA. Beck Spring Dolomite sampling localities: Saratoga Springs, Alexander Hills and the Kingston Ranges are outlined in red. Map modified from Macdonald et al. (2013). B) Generalised stratigraphic section of the Pahrump Group in Death Valley, sample localities shown in stratigraphy (modified after Mahon et al., 2014).

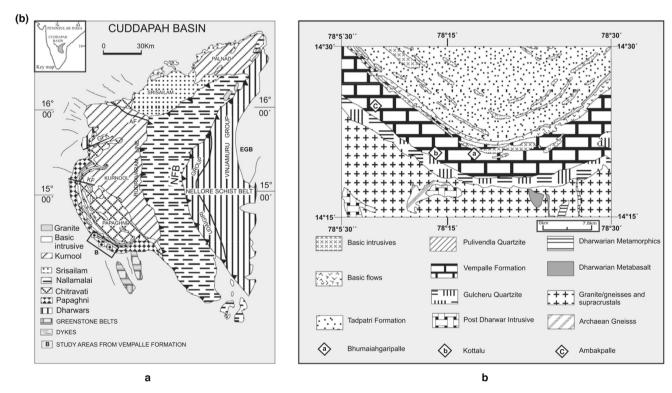


Fig. 1. Geological maps of the study area. (A) Generalized geological map of the Cuddapah Basin. (B) Detail of the southernmost Papaghni sub-basin showing locations of the measured sections investigated in this study.

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

RMetS

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

front-end interactive The 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 masterslave 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

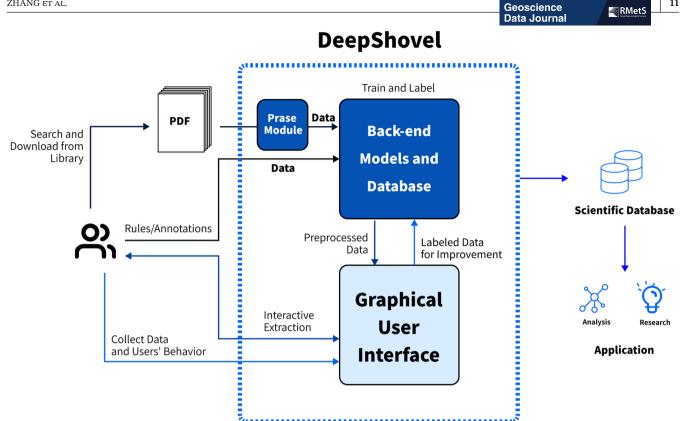


FIGURE 8 GeoDeepShovel System overview. GeoDeepShovel consists a PDF prasing module, a backend server including some artificial intelligence models and a interactive graphical user interface.

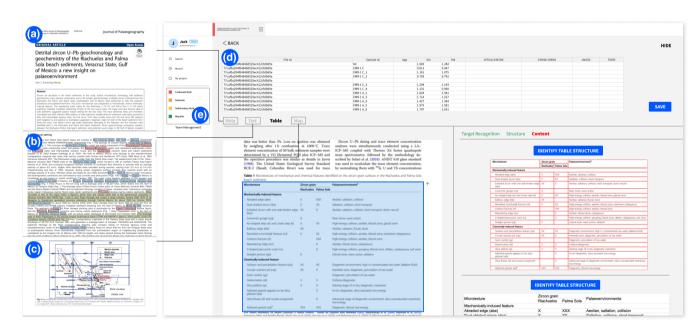


FIGURE 9 User Interface of GeoDeepShovel. The main part of this figure illustrates the table extraction and integration functions (d), while the system can also support meta information extraction (a), text extraction (b), map recognition and location extraction (c), and team and document management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

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

Data extraction process 3.3

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

11

 I2
 Geoscience

 Data Journal
 Mets

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

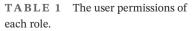
We use weak-supervision learning models and rules to help highlight the focused keywords in texts and the samples' features to help them add these words to the database. To extract academic entities from papers in the format of

Carbonate Platform Members Carbonate Platform Curbonate Platform Curb	Carbonate Platform Intervence Carbonate Platform Intervence Carbonate Platform Intervence Sk-lODP Intervence Curbonate Platform Morphological Data Estraction Intervence PETM-Continental shelf Intervence	Pr	roject M							Sedimentary F
Carbonate Platform Morphological Data Estraction	Carbonate Ratform Morphological Data Editaction	_	-]			NEW PROJECT	
SR-IODP Caribonate Platform Morphological Data Extraction	SR-IODP Caribonate Platform Morphological Data Extraction				raction					
SR-IODP Carbonale Platform Morphological Data Extraction	SR-IODP Carbonate Radrom Morphological Data Estraction									
PETM-Continental shelf	PETM-Continental shelf				raction					T Tom particle states of particle states of the states of
Continental Shell Profile	Continental Shell Profile	Ĩ	PETM-C	continental shelf						
			Continenta	ul Shelf Profile						
			continenta							
			continenta							
			Contraction							
Dolomite () F12 Execut		Do		3						F12 скаронт
Dolomite (9) F12 EXPORT Q (2) Axis (2) Axis (2) Axis			olomite 4	5		∀ •000				F12 EXPORT
Q V risk B Ms. Respect Unstables Landbles Landbles Landbles Mail Tapes 1 Mail Tap	Na. Minitare 1 Macqui Land Glazz Later Upder 1 Inspective Importine 1 Meta 1 Riper 0 Table 1		olomite 4	He Kurse 0	Masgar	J	Leveriptes 1	Ingelitier	kggt fire 1	
Q ♥ was	Ins. Finitery I Analytic LaseEday LaseEday LaseEday Impactive Impact	0	olomite (FleXane 0 Spectragophic Determination of CaO and Mr0 is Linescone and Detarriles (SPLDG)		Latert Editor				Neta O Figure O Table O
Q V N0 In Notematic Condent Scongraphic Intermeter Condent Level Status Level Status Level Status Level Status Level Status Level Status Level Status Level Status Level Status	In Material Anaget Land Material Land Material <thland materia<="" th=""> <thland material<="" th=""> <thland< td=""><td></td><td>Slomite &</td><td>File Name 0 Spectrappipt Determination of CaD and MgD is Unrestone and Debanish: GEOLOG CAL NOTS Diagnesis of a related Velocitation Versportist sequence of the Middle Macchickle Middle Finissick, the extrana Cooper Resp. NE 5</td><td>take charge</td><td>Latert Editor Jack</td><td>2821.07.27 14:55</td><td>laora(323</td><td>2821.06.11 14:25</td><td>Nets 0 Figure 0 Table 0 X X</td></thland<></thland></thland>		Slomite &	File Name 0 Spectrappipt Determination of CaD and MgD is Unrestone and Debanish: GEOLOG CAL NOTS Diagnesis of a related Velocitation Versportist sequence of the Middle Macchickle Middle Finissick, the extrana Cooper Resp. NE 5	take charge	Latert Editor Jack	2821.07.27 14:55	laora(323	2821.06.11 14:25	Nets 0 Figure 0 Table 0 X X

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.

Q				A sure							
	No.	File Name 0	Principal	Latest Editor	Latest Update 0	Import User	Import Time 0	Neta 0	Figure 0	Table 0	Text
0	1	Spectrographic Determination of CaO and MgO in Limestone and Dolomite1: GEOLOGI CAL NOTES	tako charge	Jack	2821.07.27 14:55	lasrail23	2021.06.11 14:25	×	×		
	2	Diagenesis of a mixed alliciclastic/evaporitic sequence of the Middle Maschelkalk (Middl e Triassic), the Catalan Coastal Range, NE S pain	Tom	Tom	2821.07.28 11.04	laerail23	2021.06.11 14:21	×	×		
0	3	Porosity Styles of the Midale Field in Willisto n Rasin of Southeastern Saskatchewan: ABS TRACT	Tom	Tom	2821.07.02 22:29	laterai323	2821.06.11 14:28	×	×		×
	+	Stable isotopic and elemental characteristic s of recent tata from a karatic Kika River (so with - cost Slovenia); useful environmental p roxies?	David	David	2821.06.29 13.26	laerail23	2021.06.11 14:19	×	×		×
0	5	Cause of dominance by sheetflood vs. debri 5 - flow processes on two adjoining alluvial fare, Death Valley, California	Tom	Tom	2821.06.27 20:15	laterai323	2821.06.11 14:17	×	×		×
	6	Paleozoic Rocks of North-Central Nevadal			2821.05.11 14:16	laorai123	2021.06.11 14:16	×	×	×	×
	7	Delomite Houritains and the origin of the d olicenite tock of which they mainly comite historical developments and new perspecti- ves	take charge	Willam	2821.07.27 17:17	laerai123	2021.06.11 14:15	×	×		
0		Litholecies and cyclicity of Hohilla evaporit e basins on the rifted margin of the Levant i in the Late Triassic, Makintesh Ramon, south em Israel	take charge	Tom	2822.01.06.22:35	luoruit23	2021.06.11 14:13	×	×		
0	9	Genesis of Mississippi Valley-Type Load and Zinc One Deposits and Consequent Explorat ion Thinking: ABSTRACT	Withdow	Jack	2021.05.29 14:19	laorai123	2021.06.11 14:12	×	×		×
٥	10	Origin of phosphorite of the Late Precambri an Gangolihat Dolomites of Pithoragarh, Ka meun Himalaya, India	tako chorge	David	2021.05.25 14:33	laorai123	2021.06.11 14:11	×	×		×
				< 🖸	234 >						

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



Cwelow po	onan index accounts into the second sec		
Q. Search	<back Meta Text Table Map</back 		F2 INTEGRATE
Recent My project Carbonate Rack Dolomite	Amstrong Aftein Journal of Palaeogeography https://doi.org/11.1186/43201-632-00005-9	Title	EXPROT
Setämentary Rock	ORIGINAL ARTICLE Open Access Detrital zircon U–Pb geochronology and geochemistry of the Riachuelos and Palma	Abstract	Zircons are abundant in the beach sediments. In this study, surface microtexture, mineralogy, bulk sediment geochemistry, trace element composition and U-Pb isologic geochronology of detribut across collected from the Richardos and Planet Sob beach areas, southwestern Gulf of Mexico were performed to infer the sediment provenance and paleeomivorment. The aircom microtexture was categorized as mechanical-ward/or chemicalitylocide faitures. The weathering index values for the Riachaulos (-72–77) and Palma Sola (-71–74) beach
	Sola beach sediments, Veracruz State, Gulf of Mexico: a new insight on	Year	2020
	palaeoenvironment John S. Armstrong Altrino	Author	John S Armstrong-Altrin
	Abstract	Journal	Journal of Palaeogeography
	Zircons are abundant in the backh sediments. In this study, surface microtexture, mineralogy, bulk sediment geochemistry, trace element composition and UP-b isotopic geochemology of detribuit Zoross collected from the Rischuelos and Palma Sola backh areas, southwestern Gulf of Mexico were performed to Infer the sediment provemace and palaeemovimonent. The zircon microtexture was categorized as mechanically-and/or chemically-	ISSN	25244507
	induced features. The weathering index values for the Reachuelos (~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 strace elemental ratios and chondrite- normalized are earch element (REE) patteres to plak sediments revealed that the sediments were likely sourced by	Volume	9
	Feixi and intermediate igneous rocks. And the zircon TNU ratios (mostly more than 0.2) and zircon FEE patterns (with negative E us and positive C as anomalies upgested a magnatic origin for both of the beach sediments from these two areas. Two distinct zircon age peaks respectively beforinging to the Paleosocia and the Cenosoic were identified both in the Rischuelos and Palma Sob beach sediments. Zircon geochemology comparison research	Issue	٥
	between the Rischuelos-Palma Sola beach sediments and potential source areas in SW Galf of Mexico revealed that the source terrane supplied the Paleozoic zircons of this study was identified as the Mesa Central Province (MCP), and the Cenzoic zircons were transported from the nearby Eastern Alkaline Province (EAP). Moreover,	Page	Start 1 End 27
	although the Precambrian ziccons were very few in the studied sediments, their geochronology and geochemistry irealist still could infer that they were contributed by the source teranes of Grenvillian igneous suites in the Quadca and the Chiapas Massif Complexes. Known of Detailed interaction and the Chiapas Detailed and the Chiapas the Detailed and the Chiapas the Chiapas and the Detailed and the Chiapas the Detailed and th	DOI	10.1186/s42501-020-00075-9
	Keywords: Detritil zircon, Bach sediment, U-Po dating, Zircon grain morphology, Microtexture, Mineralogy, Geochemisty, Geochronology, Gulf of Mexico	Link	https://journalofpalaeogeography.springeropen.com/articles/10.1186/s42501-020-00075-9
		Publisher	SpringerOpen
	Consepondence: amistiong@cmarl.unam.me; john_ami@yahoo.com Unidad de Procesos Oceánicos y Costeros, Instituto de Ciencias del Mar y	Language	English

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.

	temportuni nun en	
J Jack Oww jaddawymaa.cs	<back Meta Text Table Map</back 	INTEGRATE
C. Search C. Search C. Myproject C. Myproject C. Grafoxult Ruck C. Grafoxult Ruck C. Grafoxult Ruck C. Stellerstay Ruck C. Hypolite	Abstract Abstract	EXPLOYED THE DESCRIPTION OF THE
	<text><text><text><text></text></text></text></text>	average velocity of 6 cm/s. Morreal-Graze and Salas de Lan (1990 document@g) and [Salar 07 Mosice water circulation in and hydrodynamic contributions are controlled by loce current and anticipcionic rings. The climits of field of Mosice are control means of the control of the control of the climits of the
	Amstrong-Altrin Journal of Policeogeography (2020) 928 Page 2 of 27	metasedimentary rocks of the (6) in the second s
	1 Introduction Zironn is a common accessory mineral in clastic action ments, which retrying and transport; differences in ziron age populations in sediments are linked to the nature age data and trace detiment data is nor enable to heter	In total, 35 bulk sediment samples (2 kg each were collected from the Riachuelos frumber of samples n 15 and Palma Sola (n 20 bash areas, where the waves reach the coast during high tids. The mineralogy of 10 sediment samples in the status of solarity in the sediment of the sediment of the sediment of the sediment of solarity. Universidal Naclanda Automa de Malco (UMAX - Flag validation (XR) was generated with an accelerating voltage of 40 kV and a filament current of 30 mA, using CuK radiation and a graphite monochromator.

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.

13

Geoso Data J	lournal	RMetS											
	CHRIST SOTONS IN LUSTIC ROOM NEOPICTEROIDE CHRIST CTLL	MOTHE X											
Jack www	< васк												INTEG
earch	Meta	Text Table Ma	ар										INTEG
ecent						Target Rec	ognition Structure Co	ontent					
y project	F5	element concentration of 30 bulk se	ediment	00 °C. Trace samples were	Zircon U-Pb dating and trace element concentration analyses were simultaneously cond ICP-MS coupled with Thermo Xii \oplus Move mass spectrometry, followed by th	F6		IDEN.	TIFY TA	BLE STRUC	TURE	\$ ←	→ Adjı Merge
bonate Rock		the operation procedure was simil	lar as de	tails in Jarvis	scribed by Solari et al. (2018). ANIS		Microtexture	Zircon gr		Relacionaria		~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~	< -
te		BCR-2 (Basalt, Columbia River)	was us	ed for trace	was used to recalculate the trace ele by normalizing them with ²⁹ Si. U an ↑ Direction		Microtexture Mechanically-induced feature		s Palma Sol		τ ⁻	Θ	Add/D
ntary Rock		Table 1 Microtextures of mechanical an	nd chemi	al features iden	tified on the zircon grain surfaces in the Riac		Abraded edge (abe)	х	XXX	Aeolian, saltation, ci	olision		
		beach sediments	Zircon	araio	Palaeoenvironment ^a	1	Dual striated zircon (dsz) Euhedial zircon with one side broken edo	X	XX	Saltation, collision, s	short transport ollision, short transport, storm record		
				los Palma Sola			(bez)	, IX	×				
		Mechanically-induced feature					Crescentic gouge (crg) Arc-shaped step (as) and Linear step (b)	X	xx	Near shore, wave as High-energy collision	ction In, aeolian, littoral zone, glacial zone		
		Abraded edge (abe) Dual striated zircon (dsr)	X	XXXX	Aeolian, saltation, collision Saltation, collision, short transport		Bulbous edge (ble)	XX		Aeolian, saltation, fl	uvial, dune		
		Euhedral zircon with one side broken edge	• XX	x	Aeolian, saltation, collision, short transport.		Reworked conchoidal fracture (rd) Collision fracture (cf)	х	XX XXX		in, aeolian, littoral zone, nearshore subaqueo in, aeolian, littoral zone	us	
		(bez)					Meandering ridge (m/)		X	Aeolan, littoral dun			
		Crescentic gouge (crg) Arc-shaped step (as) and Linear step (ls)	×	xx	Near shore, wave action High-energy collision, aeolian, littoral zone, glacial zone		V-shaped percussion crack (vs)		х		in, gouging, littoral zone, deltaic, subaqueou	s, surf zone	
		Bulbous edge (ble)	xx	~~	Aeolian, saltation, fluvial, dune		Straight groove (sgt) Chemically-induced feature	х		Littoral zone, wave	action, saltation		
		Reworked conchoidal fracture (rcf)	x	XX	High-energy collision, aeolian, littoral zone, nearshore subaqueous		Solution and precipitation feature (s/p)	XX	XX	Diagenetic environn	ment, high in contaminated sea water (alkali	ne fluid)	
		Collision fracture (cf)		XXXX	High-energy collision, aeolian, littoral zone		Circular solution pit (csp) Grain cavities (ac)	XX	x	Intertidal zone, diag Diagenetic, percolat	enetic, percolation of sea water		
		Meandering ridge (mr)		×	Aeolian, littoral dune, subaqueous		Delamination (dl)	x	х	Collision/diagenetic	CON OF SEA WAREN		
		V-shaped percussion crack (vs) Straight groove (sgr)	×	x	High-energy collision, gouging, littoral zone, deltaic, subaqueous, surf zone Littoral zone, wave action, saltation		Silica pellicle (sp)	х	х		stu diagenetic, nearshore		
		Chemically-induced feature	^		Listen anny many action, aasanni		Adhered particle appears to be silica globule (ads)		×		ilica saturated, low-energy		
		Solution and precipitation feature (s/p)	XX	XX	Diagenetic environment, high in contaminated sea water (alkaline fluid)		Silica flower (sf) and crystal overgrowth		×	Advanced stage of low-energy	diagenetic environment, silica oversaturated	nearshore,	
		Circular solution pit (csp)	XX	х	Intertidal zone, diagenetic, percolation of sea water		Adhered particle (ad) ³	3000	XXX	Diagenetic, littoral, l	low-energy		
		Grain cavities (gc) Delamination (dl)	×		Diagenetic, percolation of sea water Collision/diagenetic								
		Silica pellicle (sp)	x	x	Starting stage of in-situ diagenetic, nearshore		_						
		Adhered particle appears to be silica globule (ads)		×	In-situ diagenetic, silica saturated, low-energy			DENTI	FY TAB	LE STRUCT	URE		
		Silica flower (sf) and crystal overgrowth		х	Advanced stage of diagenetic environment, silica oversaturated nearshore, low-energy			Zircon	arala				
		Adhered particle (ad) ^b	XOOX	3000	Diagenetic, littoral, low-energy	Microtexture	•			alma Sola	Palaeoenvironmenta		
		Ann means Abundant; AA means Common; A Armstrong-Altrin and Natalhy-Pineda (2014), Vos	et al. (2014	ent; stands for d), Hossain et al. (2)	tations after Mahaney (2002), Mathavaraju et al. (2009), Mahaney et al. (2012), (20), and Mohammad et al. (2020); ^b Adhered particles are defined as mechanical/		y-induced feature						
		chemical origin in some studies (e.g. Madhavaraj	u et al. 200	; Vos et al. 2014). F	eterring Figs. 6 and 7 for SEM images	Abraded ed	ge (abe) d zircon (dsz)	X X			Aeolian, saltation, collision Saltation, collision, short tra		
							con with one side broken edge		x		Aeolian, saltation, collision,		ansport, st
						Crescentic	jouge (crg)	х		1	Near shore, wave action		
						Arc-shaped Bulbous ed		X XX	×		High-energy collision, aeolia Aeolian, saltation, fluvial, du		al zone, gl
							onchoidal fracture (rcf)	x	×		High-energy collision, aeoli		al zone, nr
		Armstrong-Altrin Journal of Palaeogeograph	ty (2	020) 9:28	Page 8 of 27	Collision fra	cture (cf)		×	XX I	High-energy collision, aeolia	an, littora	
						Meandering V shaped p			×		Aeolian, littoral dune, subac		
						V-snaped p Straight gro	ercussion crack (vs)	x	^		High-energy collision, goug Littoral zone, wave action, s		
		are calculated by employing an ext	ornal etc	ndard zircon	uncertainty, > 20% discordance, or > 5% reverse discord-	Chemically-	induced feature				,		
		are calculated by employing an ext	207 pL	/206 ph antion	ance were not considered for interpretation. The U-Pb	Solution and	precipitation feature (s/p)	XX	х	X I	Diagenetic environment, hig		ntaminated

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

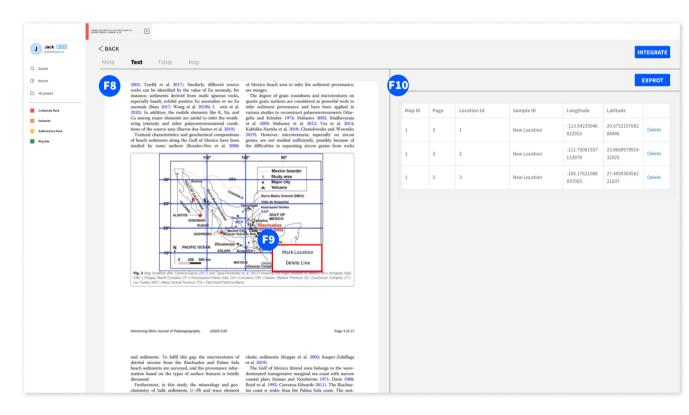


FIGURE 14 UI of Map Recognition and Location Extraction. User can adjust the recognized longitude and latitude line to get a correct coordinate range.

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 15

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

(a) **Project Settings** Table Fusion Text label Full Text era formation Full Text region Full Text subject Full Text thickness Full Text stratum Full Text age Full Text Full Text location Cancel

	Project	Settings
Text	Table Fusion	
	Keywords Group	Description
	SrNd	Strontium neodymium isotope
0	nju	detrital component
	e-IODP	International Ocean Discovery Program
	mineral-resources	knowledge graph and intelligent prediction
	Magmatite	Magmatite
	nannofossil	nannofossil
	Early Palaeozoic Marine Biodiversity	Collecting global stratigraphical data
	ELIP-UYCB	upper Yangtze cratonic basin S2S system
		Cancel Confirm

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

Geoscie Data Jou		5												ZH
_		-												
	CHIEREN DOTOPES IN CLUSTIC RODOS AND THE KOMPOTENCED CHIEREN CYCLE													
ack over	< васк													
n 🕻	FileId			Sample Id	Age	Sm	N	8	147Sm/144/Nd	143Nd/14	INd	εNd((0) TDM2	
nt	af0a294f648468529ec412cfd0d9e 77caf0a294f648468529ec412cfd0d9e			Mc EMR-1 C		1.869 0.813	1.284							
	77caf0a294f648468529ec412cfd0d9e			EMR-1 C_1		1.161	1.075							
oject	77caf0a294f648468529ec412cfd0d9e			EMR-1 C_2		0.793	0.792							
	77caf0a294f648468529ec412cfd0d9e 77caf0a294f648468529ec412cfd0d9e			EMR-1 C_3		1.294	1.193							
	77caf0a294f648468529ec412cfd0d9e			EMR-1 C_4		1.131	0.958							
nate Rock	77caf0a294f648468529ec412cfd0d9e			EMR-1 A_2		1.824	1.581							
ite	77caf0a294f648468529ec412cfd0d9e 77caf0a294f648468529ec412cfd0d9e			EMR-1 A_3 EMR-1 A_4		1.714	1.527							
	77caf0a294f648468529ec412cfd0d9e			EMR-1A 5		1.975	1.581							
ientary Rock	77caf0a294f648468529ec412cfd0d9e			EMR-1A_6		1.797	1.651							
	data was better than 5%. Loss on			d Zircon U-Pb dating and trace ele				Target Ree	cognition Structure Co	ontent				
	by weighing after 1 h combusti element concentration of 30 bulk determined by a VG Elemental P	sediment sa PQII plus I	amples wer CP-MS an	re ICP-MS coupled with Thermo Xi mass spectrometry, followed by the	ii Series quadrup he methodology	ole de-				IDENT	IFY TAE	LE STRUC	TURE	
	element concentration of 30 bulk determined by a VG Elemental P the operation procedure was sim (1988). The United States Geole	sediment se PQII plus I nilar as deta ogical Surv	amples wer CP–MS an ails in Jarvi ey Standar	re ICP–MS coupled with Thermo Xi mass spectrometry, followed by th scribed by Solari et al. (2018). ANIS was used to recalculate the trace ele	ii Series quadrup he methodology T 610 glass stand ement concentration	ole de- ard on,			Microtexture	Zircon gr		Palaeoenvironme		
	element concentration of 30 bulk determined by a VG Elemental P the operation procedure was sim (1988). The United States Geole BCR-2 (Basalt, Columbia River	sediment sa PQII plus I nilar as deta ogical Surve r) was use	amples wer CP–MS an ails in Jarvi ey Standar d for trac	re ICP-MS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). ANIS was used to recalculate the trace ele by normalizing them with ²⁹ Si. U and	ii Series quadrup he methodology T 610 glass stand ement concentrat ad Th concentrati	ole de- ard on, ons			Microtexture Mechanically-induced feature Associed edge (abs)	Zircon gr	sin	Palaeoenvironme	nt ^e	
	element concentration of 30 bulk determined by a VG Elemental P the operation procedure was sim (1988). The United States Geole BCR-2 (Basalt, Columbia River) Table 1 Mcotestures of mechanical back sedments	sediment si PQII plus I nilar as deta ogical Survo r) was used and chemica	amples wer CP–MS an ails in Jarvi rey Standar d for trac al features ide	ICP-MS coupled with Thermo Xi ans spectrometry, followed by this scribed by Solari et al. (2018). ANIS was used to recalculate the trace eleve by normalizing them with ²⁹ Si. U an entified on the zircon grain surfaces in the Riac	ii Series quadrup he methodology T 610 glass stand ement concentrat ad Th concentrati	ole de- ard on, ons			Mechanically-induced feature Abradied edge (abe) Dual striated aircon (dsd)	Zircon gr Riachueld X X	sin s Palma Sol	Aeolian, saliation, c	nt* collision short transport	
	element concentration of 30 bulk determined by a VG Elemental P the operation procedure was sim (1988). The United States Geol BCR-2 (Basalt, Columbia River; Table 1 Microtestures of mechanical	sediment so PQII plus I0 nilar as deta ogical Survo) was used and chemica Zircon gr	amples wer CP-MS an ails in Jarvi ey Standar d for trac al features ide rain	e ICP-MS coupled with Thermo Xi mass spectrometry, followed by this scribed by Solari et al. (2018). ANIS diversified on recalculate the trace by normalizing them with ²⁰ SL U an entified on the zircon grain surfaces in the Rise Palaecenvironment ⁴	ii Series quadrup he methodology T 610 glass stand ement concentrat ad Th concentrati	ole de- ard on, ons			Mechanically-induced feature Abraded edge (abe)	Zircon gr Riachueld X X	sin s Palma Sol	Aeolian, saliation, c	nt ^a	
	element concentration of 30 bulk determined by a VG Elemental P the operation procedure was sim (1988). The United States Geodo BCR-2 (Basalt, Columbia River) Table 1 Microtestures of mechanical backs sedments: Microtesture	sediment so PQII plus I0 nilar as deta ogical Survo) was used and chemica Zircon gr	amples wer CP–MS an ails in Jarvi rey Standar d for trac al features ide	e ICP-MS coupled with Thermo Xi mass spectrometry, followed by this scribed by Solari et al. (2018). ANIS diversified on recalculate the trace by normalizing them with ²⁰ SL U an entified on the zircon grain surfaces in the Rise Palaecenvironment ⁴	ii Series quadrup he methodology T 610 glass stand ement concentrat ad Th concentrati	ole de- ard on, ons			Mechanically-induced feature Abradiel edge (abe) Dual swited atron (dsc) Euhedgal atron with one side braken ed (bsc) Crescentic gruge (ctg)	Zircon gr Riachueld X X	sin s Palma Sol	Aeolian, saliation, collision, Aeolian, saliation, collision, Aeolian, saliation, collision, Near shore, wave a	nt [*] utiliaion ubat transport. utiliaios, short transport, storm record utiliaios.	
	element concentration of 30 bulk determined by a VC Elemental P the operation procedure was im (1998). The United States Grook ECR-2 (Basil, Columbia River, Table 1 Motoretours or mechanical basch address: Microtestane Microtestane	sediment si PQII plus I0 nilar as deta ogical Survo r) was user and chemica Zircon gr Riachuek	amples wer CP-MS an- ails in Jarvi ey Standar- d for trac al features ide rain os Palma Sol	e ICP-MS coupled with Thermo XJ mass spectrometry, followed by the scribed by Solari et al. (2018). ANSI was used to recalculate the trace else by normalizing them with ²⁸ SL U an entitled on the zeroon grain surfaces in the Riac Palaesenvironment [*] la	ii Series quadrup he methodology T 610 glass stand ement concentrat ad Th concentrati	ole de- ard on, ons			Mechanically induced feature Astrodect edge (abe) Dual strated encore (adu Euhaded) attoom with one side broken et (bio) Cresentic gouge (og) Arc shaped step (ad) and Liver step (b)	Zircon gr Riachueld X X	sin s Palma Sol	Aeolan, salation, o Salation, collision, Aeolan, salation, o Near shore, wave a High-energy collisi	nt ⁴ collision short transport collision, block transport, storm record colon colon exection. Taxoni zone, glacul zone	
	element concentration of 30 built determined by a VC Elemental P the operation procedure was im PCP-2 (Insult, Catambia River Table 1 Monotonies of mechanical School and School (Insult) Microtenter Rechanical) induced frame Rechanically induced frame	sediment sc PQII plus IA ailar as deta ogical Survo) was user and chemica Zircon gr Riachuek	amples wer CP-MS an ails in Jarvi ey Standard for trac al features ide rain os Palma Sol XXX	e ICP-MS coupled with Thermo Xi and mass spectrometry. followed by the scribed by Solari et al. (2018). ANISI was used to recalculate the trace ele by normalizing them with ²⁷ Si. U an entitled on the zircon gain surfaces in the Bias Plaaseemiciennem ⁴ is realistic additional and the second	ii Series quadrup he methodology T 610 glass stand ement concentrat ad Th concentrati	ole de- ard on, ons			Mechanically-Induced feature Annoded edge (ale) Dual strianed arcon (alcd) Euhodial arcon with one side broken ed (ber) Crecomic goage (cg) Arc-shaped step (as) and Linear step (b) Bulloou edge (ble)	Zircon gr Riachueld X X	sin s Palma Sol	Arcolan, saltation, o Saltation, collision, Aeolan, saltation, collision, Aeolan, saltation, d Near shore, wave a High-energy collisi Aeolan, saltation, f	mt collsion short transport sollins, short transport, storm record ectors or, aveiding	
	element concentration of 30 built determined by a VC Elemental P the operation procedure was imin (1998). The United States Grook RCR-2 (Basilt, Columbia Rivery Table 1 Mortosteurs on mechanical basics and mechanical basics and the state of the state Microsteurs Microsteurs Microsteurs Microsteurs that basic and the state of the state basic and the state of the state basic and the state of the state of the state basic and the state of the state of the state basic and the state of the state of the state of the state basic and the state of the state of the state of the state basic and the state of the state of the state of the state of the state basic and the state of the	sediment sc PQII plus I0 hilar as deta ogical Survo) was user and chemica Zircon gr Riachuek X X	amples wer CP-MS an ails in Jarvi ey Standar d for trac il features ide rain os Palma Sol XXX	ECP-ASS coupled with Thermo Xi mass spectrometry, followed by th scribed by Solari et al. (2018), ANIS was used to recalculate the trace else by normalizing them with. ³⁵ SL U an enfield on the aircon quark scales in the Rise Palascentroment. ⁴ Andra, substain, collision Substain, collision, short temport	ii Series quadrup he methodology T 610 glass stand ement concentrati ad Th concentrati chuelos and Palma	ole de- ard on, ons			Mechanically induced feature Astrodect edge (abe) Dual strated encore (adu Euhaded) attoom with one side broken et (bio) Cresentic gouge (og) Arc shaped step (ad) and Liver step (b)	Zircon gr Riachueld X X	sin s Palma Sol	Palaeoenvironme Aeolan, salation, co Satation, collision, Aeolan, salation, c Near shore, www a High-energy collisi High-energy collisi	nt ⁴ collision short transport collision, block transport, storm record colon colon exection. Taxoni zone, glacul zone	
	element concentration of 30 built determined by a VC Elemental P the operation procedure was im PCP-2 (Insult, Catambia River Table 1 Monotonies of mechanical School and School (Insult) Microtenter Rechanical) induced frame Rechanically induced frame	sediment sc PQII plus I0 hilar as deta ogical Survo) was user and chemica Zircon gr Riachuek X X	amples wer CP-MS an ails in Jarvi ey Standard for trac al features ide rain os Palma Sol XXX	e ICP-MS coupled with Thermo Xi and mass spectrometry. followed by the scribed by Solari et al. (2018). ANISI was used to recalculate the trace ele by normalizing them with ²⁷ Si. U an entitled on the zircon gain surfaces in the Bias Plaaseemiciennem ⁴ is realistic additional and the second	ii Series quadrup he methodology T 610 glass stand ement concentrati ad Th concentrati chuelos and Palma	ole de- ard on, ons			Mechanically induced feature Abaded edge (abit) Dual proved arcon (ibd) Exherbal accon who ne add braken ed bad) Crossortine goage (org) Arc-shaped step (abit and Linear step (ib Buckon edge (Bit) Buckon edge (Bit)	Zircon gr Riachueld X X	sin s Palma Sol	Palaeoenvironme Aeolan, salation, co Satation, collision, Aeolan, salation, c Near shore, www a High-energy collisi High-energy collisi	mt zalatom Jahot streppot schor streppot schor kontos jone jakal prore kundu dare on kenklik filtad jahon kenabaseus kundu dare	
	element concentration of 30 built determined by a VC Elemental P the operation procedure was im (1916). The United States Gook BCR 2 (basils, Columbia Browy Table 1 Montestores of mediated at teach indexets Montestore	sediment sc PQII plus IA hilar as deta ogical Survo) was user and chemica Zircon gr Riachuek X X	amples wer CP-MS an ails in Jarvi ey Standar d for trac il features ide rain os Palma Sol XXX	ECP-ASS coupled with Thermo Xi mass spectrometry, followed by th scribed by Solari et al. (2018), ANIS was used to recalculate the trace else by normalizing them with. ³⁵ SL U an enfield on the aircon quark scales in the Rise Palascentroment. ⁴ Andra, substain, collision Substain, collision, short temport	ii Series quadrup he methodology T 610 glass stand ement concentrati ad Th concentrati chuelos and Palma	ole de- ard on, ons			Michaeled stay later Attuated stay later Coal mined acros that Lacked lack with the safe toden of Coronner gauge ring Are-dupped top bial and there may full blacker skip blaf Revoluted conclusion full Collaion historie (i) Manadering rings (iii) Varkeed genesisment ack kit	Zircon gr Riachueld X X	sin s Palma Sol	Palaecervironme Aeolan, saltation, c Sattation, collisien, Aeolan, saltation, c Near shore, www a High-energy collisi High-energy collisi High-energy collisi High-energy collisi	et and the temport collision for temport collision for temport collision for temport collision for temport con addition (though more reactions teleparation on addition (though more reactions teleparation on addition (though more reactions teleparation on addition (though more reactions teleparation) on addition (though more reactions) teleparation (though more reaction) teleparation (though more rea	one
	element concentration of 30 built determined by a VC Elemental P the operation procedure was imin (1998). The United States Grook ECR-2 (Basilt, Columbia Briver, Table 1 Motostones on meta-ancal basch andremets. Microstenier Machanically induced feature Anadored opti India Data Statest datons into Liabelad intered discons into Liabelad intered discons into Concomic groups (trig) Archanes in the and Linear strap Nel Archanes in the and Linear strap Nel Archanes in the and Linear strap Nel	sediment s: PQII plus II nilar as deta gojcal Surv r) was user and chemica Zircon gr Riachueld X X X age XX X X	amples wer CP-MS an ails in Jarvi ey Standar d for trac il features ide rain os Palma Sol XXX	ECT-ASS coupled with Thermo Xi mass spectromaries, followed by this scribed by Solari et al. (2019). ANIS was aud to recallate the trace of who maintaing them with "Ss. U an enter the solar solarity of the solarity of the the solarity of the solarity of the solarity haloesentement" Aester, satisfar, calition, Satisfar, calition, solar tamport Aester, satisfar, calition, solar tamport Aester, satisfar, calition, solar tamport, Aester, satisfar, calition, solar tamport, Aester, satisfar, calition, solar, timad ano, glaciti hybridency efficiency, solar, timad ano, glaciti hybridency efficiency, solar, timad ano, glaciti	ii Series quadrug he methodology folo glass stand ement concentrat d Th concentrat chuelos and Palma	ole de- ard on, ons			Modunically induced feature Another days latery Data strends around half Guidessid around half Guidessid around half another around another another Another around another Bibliocen adapt Bibli Bibliocen adapt Biblio Calabon Instance (B) Mandeding regist mod	Zircon gr Riachueld X X	sin s Palma Sol	Palaeoenvironme Avolan, satation, colision, Saltation, colision, Aeolan, satation, co Near shore, wave a High-energy colisis Aeolan, satation, f High-energy colisis Aeolan, satation, f High-energy colisis Aeolan, itsoral du	et and the temport collision for temport collision for temport collision for temport collision for temport con addition (though more reactions teleparation on addition (though more reactions teleparation on addition (though more reactions teleparation on addition (though more reactions teleparation) on addition (though more reactions) teleparation (though more reaction) teleparation (though more rea	one
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was im (1988). The United States Gook BCR2 (Baash, Columbia Breve) Table 1 Mortesteares of medianical teach sudences Moreteares Mortesteare Mortesteare Mortesteares Mortesteares Mortesteares Mortesteares Mortesteares Academic gray and Leader states index Bed Concents: grays (Est) Ac-shaped targe in the same bit Bit Academic gray and Learn rap bit Bit Academic gray and Learn rap bit Bit Academic grays (Est)	sediment s: PQII plus II and as deta gical Surve was used and chemica Zircon gr Riachuele X X X A age XX X X X X X	amples wer CP-MS an ails in Jarvi vy Standar, d for trac il features ide rain os Palma Sol XOX XX XX XX	KC-NAS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). AVIS was used to recalculate the tarce of the scribed by the scribed by the scribed in the scribed on the accord of the scribed factors values, collars, the scribed by the scribed factors, values, collars, the target, stem factors, values, collars, the target, stem factors, scribed, acts, factor acc, glack target, action, scriber, the scribed by the scribed scribed by the scribed by the scribed by the scribed scribed by the scribed by the scribed by the scribed by the scribed scribed by the scribed by	ii Series quadrug he methodology T 610 glass stand ement concentrati d Th concentrati chuelos and Palma	ole de- ard on, ons			Mechanically induced feature Assolid objy lidet Onal model all aroms table Exclusion with one state backen of back Compared and the state of the state of the Analysis of the state of the state of the Researched conclusion in the Contens in these in the Mechanical project of the State of the state in the Mechanical project of the State of the state of the state of the state of the State of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of the state of the state of the state of the State of the state of the State of the state of the state of the state of the state of the State of the state of the	Zircon gr Riachueld X X	sin s Palma Sol	Palaecentrironme Aeolan, saturion, collision, Saturion, collision, Aeolan, saturion, o Near shore, wave a High-energy, collisi Aeolan, saturion, f High-energy, collisi Aeolan, Ittoral du High-energy, collisi Uttoral zone, wave	et* and	
	element concentration of 30 built determined by a VC Elemental P the operation procedure was im (1998). The United States Grook ECR-2 (Basilt, Columbia Briver, Table 1 Morotonices on mechanical basilt Morotonices on mechanical basilt metric from the Rechanically induced feature Adapted region that Lindeal atoms that Concome group lengt Archange Ited Builton events on soft basiles and and the soft feature Rechanged instruction that Lindeal atoms with one side basiles Concome group lengt Archanged Ited and Lindeal atoms plut Builtons events Rechanged Ited and Lindeal atoms plut Builtons events	sediment s: PQII plus II nilar as deta gojcal Surv r) was user and chemica Zircon gr Riachueld X X X age XX X X	amples wer CP-MS an alls in Jarvi ey Standar d for trac al features ide rain xox xox xx xx xx xx xx xx xx	ECT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2019). ANSI was used to recallate the trace of the scribed on the accors gain surfaces in the Nix Placeseniument Anders, statistics, collision, Statistics, collision, solarity, and Anders, statistics, collision, Statistics, collision, solarity targets, statistics, Statistics, collision, solarity, statistics, statistics, Anders, statistics, collision, short transport, Statistics, collision, short transport, storm, Nair show, wave action Nair show, seeds, total care, parait Anders, statistics, Add, diver Hybrengy collision, seedin, throid zone, reserving Statistics, Collision, short, Statistics, Statisti	ii Series quadrug he methodology T 610 glass stand ement concentrati d Th concentrati chuelos and Palma	ole de- ard on, ons			Modunically induced feature Another days latery Data strends around half Guidessid around half Guidessid around half another around another another Another around another Bibliocen adapt Bibli Bibliocen adapt Biblio Calabon Instance (B) Mandeding regist mod	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palaecenvironme Aeolan, saturion, colsion, Aeolan, saturion, colsion, Aeolan, saturion, colsion, Aeolan, saturion, C High-energy collisi Aeolan, saturion, High-energy collisi Aeolan, Imcol du High-energy collisi Aeolan, Imcol du High-energy collisi Aeolan, Imcol du Uttoral zone, wwe Diagenetic environ	et and the temport collision for temport collision for temport collision for temport collision for temport con addition (though more reactions teleparation on addition (though more reactions teleparation on addition (though more reactions teleparation on addition (though more reactions teleparation) on addition (though more reactions) teleparation (though more reaction) teleparation (though more rea	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was im (1988). The United States Gook BCR2 (Baash, Columbia Breve) Table 1 Mortesteares of medianical texth sudences Moreteares Moreteares Moreteares Moreteares Moreteares Academised program Lands and more than Lands and program Moreteares Academised program More and Data Market Bed Concents page (print) Ac-shaped target and Lanar ang bi Blabbac edge (Bla) Revender conclusion facure offi	sediment s: PQII plus II and as deta gical Surve was used and chemica Zircon gr Riachuele X X X A age XX X X X X X	amples wer CP-MS an alls in Jarvie y Standard d for trac d for trac al features ide rain s Palma Sol XXX XX XX XX XX XX XX XX XX XX XX XX	ECT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). AVIS was used to recalculate the tarce of the scribed by the scribed by the scribed in the scribed on the accord on the face of actions, unlines, short temport Actions, subtrant, collison, short temport Actions, subtrant, collison, short temport, stem Rev flow, week scribed accord, scribed accord, scribed reprinterge collison, action, short temport, stem Rev flow, seeking, that core, parts reprintergy collison, action, scribed accord, scribed reprintergy collison, action, fund, done reprintergy collison, action, fund, done reprintergy collison, action, fund, done reprintergy collison, action, fund, done reprintergy collison, action, fund, done	ii Series quadrug he methodology T 610 glass stand ement concentrati d Th concentrati chuelos and Palma	ole de- ard on, ons			McConstanting induced feature Annobel edge (alex) Deal entered around table Constanting and tables of Constanting any strip And edged tables (alex) and Resoluted (antiched) factors are 0.0 Contourned and tables of Antiped (antiched) factors and 0.0 Stragged (antiched) factors Stragged (antiched)	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palaecenvironme Aeolan, saturion, colsion, Aeolan, saturion, colsion, Aeolan, saturion, colsion, Aeolan, saturion, C High-energy collisi Aeolan, saturion, High-energy collisi Aeolan, Imcol du High-energy collisi Aeolan, Imcol du High-energy collisi Aeolan, Imcol du Uttoral zone, wwe Diagenetic environ	ere states And stronger, there would be also and the stronger, there would be the stronger that and the stronger that the stronger that and the stronger that the stronger that and the stronger that the stronger that the stronger that and the stronger that the stronger the stronger the stronger that the stronger th	
	elementi concentration of 30 bulk determined by a VC Elemental P the operation procedure was imp PCP-32 (Basel). Contamine Rever Table 1 Monostroner of mechanical is reached and the second second second Microsteners Microsteners Microsteners Microsteners Analet degis india India it more with one side beaten et Concert capator (Sing) Accelegation (Sing) Reversed contrologification (Sing)	sediment s: PQII plus II and as deta gical Surve was used and chemica Zircon gr Riachuele X X X A age XX X X X X X	amples wer CP-MS an alls in Jarvie y Standard d for trac il features ide rain os Palma Sol xox xx xx xx xx xx xx xx xx xx xx xx xx	ECT-ASS coupled with Thermo Xi mass spectromary, followed by its scribed by Solari et al. (2018). ANS scribed by Solari et al. (2018). Solari et al. (2018). ANS scribed by Solari et al. (2018). Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Solari et al. (2018).	iii Series quadrug he methodology TT 610 glass stand menent concentrati d Th concentrati chuelos and Palma hereord al zone hore subaqueous	ole de- ard on, ons iola			Rechards indicated future Rechards registering Data meet around the Balance and the and the data for the Balance and the and the and the Rechards the set of all classes and Rechards the set of all classes and R	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palacentrironme Acolar, salation, colsion, Salation, colsion, Acolar, salation, colsion, Acolar, salation, C. Highereng, collis Acolar, Magneregy, collis Acolar, Magneregy, collis Highereng, collis Highereng, collis Highereng, collis Highereng, collis Diagnetic environ Diagnetic environ Diagnetic environ	er	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was im (1988). The United States Gook ECR2 (Baash, Columbia Breve) Table 1 Mortestrates of medianical texth sudences Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Moretexane Mor	sediment s: PQII plus II and as deta gical Surve was used and chemica Zircon gr Riachuele X X X A age XX X X X X X	amples wer CP-MS an alls in Jarvie y Standard d for trac d for trac al features ide rain s Palma Sol XXX XX XX XX XX XX XX XX XX XX XX XX	KCT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). AVIS was need to recalculate the targe of the scribed on the accord with the scribed and the scribed on the accord with the scribed address, unlines, what temport Audies, subtant, collison, what temport text has, see a scribed and the scription of the scription, see a scription of the high-temport collison, see and the scription address, text has a scription of the scriptio	iii Series quadrug he methodology TT 610 glass stand menent concentrati d Th concentrati chuelos and Palma hereord al zone hore subaqueous	ole de- ard on, ons iola			Reduction of photocel future Reduction of photocel future Reduction of photocel future Reduction of the Reduction of the Reduction Data and a second and the Reduction In the Reduction of the Reduction Reduction of the Reduction of the Reduction of the Reduction Reduction of the Reduction of the Reduction of the Reduction Reduction of the Reduction of the Reduction of the Reduction Reduction of the Reduction of the Reduction of the Reduction Reduction of the Reduction of the Reduction of the Reduction Reduction of the Reduction of the Reduction of the Reduction of the Reduction Reduction of the Reduction	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palaeconvironme Acolan, salation, collion, Salation, collion, Acolan, salation, collion, Near shore, ware at High-emprophilic	when when the second the se	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was imp rCC-32 (Basilt, Columbia River) Table 1 Monoscotter of mechanical a second second second second second second Reclassically induced feature Academic of the second second second second Electronic Second Second Second Second Columbia Second Second Second Second Collision Rescue (c) Mandering right (c) V-sheapt Demension coals fold Second	sediment s: PQII plus II and as deta gical Surve was used and chemica Zircon gr Riachuele X X X A age XX X X X X X	amples wer CP-MS an alls in Jarvie y Standard d for trac il features ide rain os Palma Sol xox xx xx xx xx xx xx xx xx xx xx xx xx	ECT-ASS coupled with Thermo Xi mass spectromary, followed by its scribed by Solari et al. (2018). ANS scribed by Solari et al. (2018). Solari et al. (2018). ANS scribed by Solari et al. (2018). Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Solari et al. (2018). Answer at the scribed by Solari et al. (2018). Solari et al. (2018).	iii Series quadrug he methodology TT 610 glass stand menent concentrati d Th concentrati chuelos and Palma hereord al zone hore subaqueous	ole de- ard on, ons iola			Rechards indicated future Rechards registering Data meet around the Balance and the and the data for the Balance and the and the and the Rechards the set of all classes and Rechards the set of all classes and R	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palaeconvironme Acolan, salation, collion, Salation, collion, Acolan, salation, collion, Near shore, ware at High-emprophilic	er	
	element concentration of 30 bulk determined by a VC Elemental P the operation procedure was im (1988). The United States Gook ECR-2 (Ikush, Columbia Breve Fabel Microsteners of medianical book states of the Columbia States Microsteners Microsteners Data Internation Microsoft Data Internation Microsoft Data Internation Microsoft Backada Internation International Internation Records of control Backada Internation Backada Internation Records Control Backada Internation Backada Internation Records Control Backada Internation Backada Internation	sediment s PQII plus IR illar as deta ogical Surve was used and chemica Zircon gr Riadvueld X X X X X X X X X	amples wer CP-MS an alls in Jarvie of for trace of for trace of for trace of Palma Sol XOX XX XX XX XX XX XX XX XX XX XX XX XX	ECT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). ANS service of the scribe of the scribe of the scribe with the scribe of the scribe of the scribe of the scribe processing of the scribe of the scribe of the scribe advance, valuetore, collision deals, sclattere, collision, short temport Andres, subtrance, collision, short temport Andres, sclattere, collision, short temport Andres, sclattere, collision, short temport Andres, sclattere, collision, short temport inspin organic collision, short temport inspin organicoli collision, short temport inspin organic collision, short tempo	II Series quadrup the methodology Tr 610 glass stant ment concentration d'hannes and Palma intecord di zone hore subaqueous saic, subaqueous, sof z	ole de- ard on, siola			Rebaciant photoes future Rebaciant physical future Rebaciant physical future Rebaciant physical future Rebaciant physical future Rebaciant physical future Rebaciant physical future Rebaciant physical Rebaciant physical Rebaciant Rebaciant physical Rebaciant	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palaecontrironme Analian, salarisin, collision, Sattatori, collision, Analian, salarisin, collision, High-emary collision, High-emar	when when the second the se	
	element concentration of 30 builts determined by a Vol Elemental P the operation procedure was into the Appendix and Appendix and Appendix Concentration of the Appendix Appendix Microtenture Microtenture Microtenture Microtenture Appendix Appendix Appendix Appendix Laboration of the Appendix Appendi	sediment s PQII plus II ilar as deta ogical Survi) was uses and chemica Zecon ge Riachuele X X X X X X X X X X X X	amples wer CP-MS an alls in Jarvie y Standard d for trac il features ide rain os Palma Sol xox xx xx xx xx xx xx xx xx xx xx xx xx	ECT-ASS coupled with Thermo Xi mass spectromary, followed by its active the Solari et al. (2018). AVX5 with the Solari et al. (2018). AVX5 with the Solari et al. (2018). AVX5 with the Solari et al. (2018). AVX5 Parameterization of the Solari et al. (2018) Parameterization collision Solarion collision, solari support Solarion, collision, solaris, format anon, glace Andrea, sulation, collision, solaris, format anon, glace followerge collision, solaris, format anon, solaris followerge collision, solaris, tabation Dagenetic environment, high in correanment of Dagenetic environment, high in correanment of Solari et al. (2018). A solari environment and anone shell Dagenetic environment, high in correanment of Dagenetic environment, high in correanment of Solari et al. (2018). A solari et al. (2018). A solari et al. (2018). A solari Solari et al. (2018). A solari et al. (2018). A solari et al. (2018). A solari Solari et al. (2018). A solari et al. (2018). A solari Solari et al. (2018). A solari et al. (2018). A solari Solari et al. (2018). A solari et al. (2018). A solari Solari et al. (2018). A s	ii Series quadrug the methodology Tr 610 glass stant ment concentrati di Th concentrati chuelos and Palma i record al zone hore subaqueous, sof z aic, subaqueous, sof z aic, subaqueous, sof z	ole de- ard on, siola			Rebacical products future Rebacical registerior Due arread around de Cale arread around de Cale around around de Cale around around de Cale around de Rebacical de de de rebacical de Rebacical de de de rebacical de Rebacical de de rebacical de Rebacical de Rebacical de de rebacical de Rebacical de Rebacical de Rebacical de Rebacical de Rebacical de Rebac	Zircon gr Riachueld X X	S Palma Sol 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024 2024	Palaecentrionme Aedian, saturion, colliaion, Sataron, colliaion, Near shore, ware a High-energy colliai Aedian, saturoti colliaion, High-energy colliai Aedian, vaturoti colliaion, Aedian, vaturoti colliaion, Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Colliaionta (annumente Colliaionta) Colliaionta (annumente) Colliaionta (annumente) Aedianentes (annumentes) Aedianentes (annume	er	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was im (1988). The United States Gook ECR2 (Realt, Columbia Rever Fabel Morotextures of medianical totals and the state of the states in the state of the states in the states of the states determined by the states of the states of Book of the states Concents groups (stat) Acomption factors field Book on the state book on the Book on the state of the Book on the state of the Book on the state of the Book on the states of the Book of the states of the states of the Book of the states of the Book of the states of the states of the states of the Book of the states o	sediment s PQII plus IR illar as deta ogical Surve was used and chemica Zircon gr Riadvueld X X X X X X X X X	amples wer CP-MS an alls in Jarvie of for trace of for trace of for trace of Palma Sol XOX XX XX XX XX XX XX XX XX XX XX XX XX	KCT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). ANS served by Solari et al. (2018). ANS served the solarity of the solarity of the served solarity of the solarity of the solarity of the processing of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of the	ii Series quadrug the methodology Tr 610 glass stant ment concentrati di Th concentrati chuelos and Palma i record al zone hore subaqueous, sof z aic, subaqueous, sof z aic, subaqueous, sof z	ole de- ard on, siola			Reduction flyndraf future Reduction flyndraf future Reduction flyndraf future Reduction flyndraf future Reduction flyndraf future Reduction flyndraf future Reduction flyndraf flyndraf Reduction future R Reduction flyndraf flyndraf Reduction flyndraf flyndraf flyndraf flyndraf flyndraf Reduction flyndraf flyndraf flyndraf flyndraf flyndraf flyndraf Reduction flyndraf flyndraf flyndraf flyndraf flyndraf flyndraf Reduction flyndraf flyndraf flyndraf flyndraf flyndraf flyndraf flyndraf Reduction flyndraf flyndraf flyndraf flyndraf flyndraf flyndraf flyndraf Reduction flyndraf	Zircon gr %	Image: second	Palaecontrironme Analian, salarisin, collision, Sattatori, collision, Analian, salarisin, collision, High-emary collision, High-emar	er	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was in the concentration procedure was in the concentration of the concentration of the concentration of the concentration of the determined concentration from the determined concentration from the determined concentration from the determined concentration from the con- tension of the concentration of the concentration of the concentration of the concentration of the determined concentration from the con- dition of the concentration from the con- centration of the concentration from the con- centration of the concentration of the concentration of the concentration of the concentration of the concentration of the concentration of the concentration of the concentration of the concentration of the concentration	sediment s sediment s pOII plus ils s deta si was uses ned chemica zincon gr Riachuele x x x x x x x x x x x x x	amples wer CPP-MS am ails in Jarvi d for trace al features ide rain se Palma Sol XOX XX XX XX XX XX XX XX XX XX XX XX XX	ECT-ASS coupled with Thermo Xi mass spectromary, followed by its active the pSolari et al. (2010). AVX5 with the pSolari et al. (2010). AVX5 PSOLARIS and the pSolari et al. (2	ii Series quadrug the methodology Tr 610 glass stant ment concentrati di Th concentrati chuelos and Palma i record al zone hore subaqueous, sof z aic, subaqueous, sof z aic, subaqueous, sof z	ole de- ard on, siola			Rebacical products future Rebacical registerior Due arread around de Cale arread around de Cale around around de Cale around around de Cale around de Rebacical de de de rebacical de Rebacical de de de rebacical de Rebacical de de rebacical de Rebacical de Rebacical de de rebacical de Rebacical de Rebacical de Rebacical de Rebacical de Rebacical de Rebac	Zircon gr %	Image: second	Palaecentrionme Aedian, saturion, colliaion, Sataron, colliaion, Near shore, ware a High-energy colliai Aedian, saturoti colliaion, High-energy colliai Aedian, vaturoti colliaion, Aedian, vaturoti colliaion, Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Colliaionta (annumente Colliaionta) Colliaionta (annumente) Colliaionta (annumente) Aedianentes (annumentes) Aedianentes (annume	er	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was im (1988). The United States Gook ECR2 (Real, Columbia Rever Fabel Morostowers of medianical total states of medianical total states of medianical total states of the States Concent; page (pin) Acostopic and states of the Book edge (pin) Reverse (conclusion factors eff) Reverse (conclusion factors eff) States representations (see 160) Conclusion factors eff) Conclusion (pin)	sediment s PQII plus II illar as deta ogical Surviv) was uses and chemica Zercon gr Riachuele X X X X X X X X X X X X X	amples wer CP-MS an alls in Jarvie of for trace of for trace of for trace of Palma Sol XOX XX XX XX XX XX XX XX XX XX XX XX XX	ECT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). ANS served by Solari et al. (2018). ANS served the solarity of the solarity of the served solarity of the solarity of the solarity of the processing of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of t	ii Series quadrug the methodology Tr 610 glass stant ment concentrati di Th concentrati chuelos and Palma i record al zone hore subaqueous, sof z aic, subaqueous, sof z aic, subaqueous, sof z	ole de- ard on, siola			Account of the second sec	Zircon gr Rischueic Rischueic X <t< td=""><td>Image: Normal Section 1 Patients Section 2 Se</td><td>Palacentromme Acolar, salation of Salation collision, Acolar, salation of Acolar, salation of Acolar, salation of Acolar, salation of Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Diagenetic, periodi Collision/Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Collision Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Salation Sagaret of Higherengy collision Diagenetic environ Higherengy collision Higherengy collision H</td><td>er state state state state state state state state state state state state state state state state state state state state state state state state state</td><td></td></t<>	Image: Normal Section 1 Patients Section 2 Se	Palacentromme Acolar, salation of Salation collision, Acolar, salation of Acolar, salation of Acolar, salation of Acolar, salation of Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Diagenetic, periodi Collision/Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Collision Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Salation Sagaret of Higherengy collision Diagenetic environ Higherengy collision Higherengy collision H	er state state state state state state state state state state state state state state state state state state state state state state state state state	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was into the operation procedure was into the operation procedure was into the operation process of mechanical resolutions of the operation of the Carl State State State State State Concentrations of the operation of the Electrical atoms that Electrical atoms that Electrical atoms that Badden atoms that a state balance at Badden atoms that Badden atoms that a state balance at Badden atoms that a state balance at Carlians induced function Contain state per togging Gane counter (g) Balance atoms (g) Balance atoms (g) Balance atoms (g) Balance atoms (g)	sediment s sediment s pOII plus ils s deta si was uses ned chemica zincon gr Riachuele x x x x x x x x x x x x x	amples wer CPP-MS am ails in Jarvi d for trace s Palma Sol XXX XX XX XXX XXX XXX XXX XXX XXX XXX	EC-ASS coupled with Thermo Xi mass spectromary, followed by its acceled by Solari et al. (2010). AVX5 we could be Solari et al. (2010). AVX5 we could be solarized as a solarized as a protection of the acceleration of the acceleration protection of the acceleration of the acceleration Solarized as a solarized as a solarized as a solarized Solarized as a solarized as a solarized as a solarized as a Solarized as a solarized as a solarized as a solarized as a Solarized as a solarized as a solarized as a solarized as a Solarized as a solarized as a solarized as a solarized as a Solarized as a solarized as a solarized as a solarized as a Solarized as a solarized as a solarized as a solarized as a Callonadaparetic persistence of solarized callongenetic Callonadaparetic as a solarized as a solarized as a solarized callongenetic Callonadaparetic as a solarized callongenetic as a solarized callongenetic as a solarized callongenetic as a solarized callongenetic Callonadaparetic as a solarized callongenetic as a solariz	ii Series quadrug the methodology Tr 610 glass stant ment concentrati di Th concentrati chuelos and Palma i record al zone hore subaqueous, sof z aic, subaqueous, sof z aic, subaqueous, sof z	ole de- ard on, siola			Account of the second sec	Zircon gr Rischueic Rischueic X <t< td=""><td>Image: Normal Section 1 Patients Section 2 Se</td><td>Palaecentrionme Aedian, saturion, colliaion, Sataron, colliaion, Near shore, ware a High-energy colliai Aedian, saturoti colliaion, High-energy colliai Aedian, vaturoti colliaion, Aedian, vaturoti colliaion, Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Colliaionta (annumente Colliaionta) Colliaionta (annumente) Colliaionta (annumente) Aedianentes (annumentes) Aedianentes (annume</td><td>er state state state state state state state state state state state state state state state state state state state state state state state state state</td><td></td></t<>	Image: Normal Section 1 Patients Section 2 Se	Palaecentrionme Aedian, saturion, colliaion, Sataron, colliaion, Near shore, ware a High-energy colliai Aedian, saturoti colliaion, High-energy colliai Aedian, vaturoti colliaion, Aedian, vaturoti colliaion, Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Aedian-tenergy colliai Colliaionta (annumente Colliaionta) Colliaionta (annumente) Colliaionta (annumente) Aedianentes (annumentes) Aedianentes (annume	er state state state state state state state state state state state state state state state state state state state state state state state state state	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was min PCC-32 (Basel). Contention and the PCC-32 (Basel). Contention and the PCC-32 (Basel). Contention and the Recharacial priority indiced feature Actastical priority indiced feature Bactors and priority indiced feature Contents quarter field Bactors and priority indiced feature Solators and precisions feature big Content and priority indiced Solators and precisions feature big Content and priority indiced Solators and precisions feature big Content and priority indiced feature Solators and priority indiced feature Solators and precisions feature big Content and precisions feature big Content and p	sediment s sediment s pOII plus ils s deta si was uses ned chemica zincon gr Riachuele x x x x x x x x x x x x x	amples wer CPP-MS am ails in Jarvi d for trace al features ide rain se Palma Sol XOX XX XX XX XX XX XX XX XX XX XX XX XX	ECT-ASS coupled with Thermo Xi mass spectrometry, followed by the scribed by Solari et al. (2018). ANS served by Solari et al. (2018). ANS served the solarity of the solarity of the served solarity of the solarity of the solarity of the processing of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of the solarity of t	ii Series quadrug the methodology Tr 610 glass stant ment concentrati di Th concentrati chuelos and Palma i record al zone hore subaqueous, sof z aic, subaqueous, sof z aic, subaqueous, sof z	ole de- ard on, siola			Account of the second sec	Ziron g Ziron g Richarlo X X X X X XX X XX X X X X X X X X X X X XX X XX X X <td>NI Patrins 541 Patring Patring 541 Patrin</td> <td>Palacentromme Acolar, salation of Salation collision, Acolar, salation of Acolar, salation of Acolar, salation of Acolar, salation of Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Diagenetic, periodi Collision/Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Collision Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Salation Sagaret of Higherengy collision Diagenetic environ Higherengy collision Higherengy collision H</td> <td>er state state state state state state state state state state state state state state state state state state state state state state state state state</td> <td></td>	NI Patrins 541 Patring Patring 541 Patrin	Palacentromme Acolar, salation of Salation collision, Acolar, salation of Acolar, salation of Acolar, salation of Acolar, salation of Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Higherengy collision Diagenetic, periodi Collision/Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Collision Higgereng Salation Sagaret of Higherengy collision Diagenetic, periodi Salation Sagaret of Higherengy collision Diagenetic environ Higherengy collision Higherengy collision H	er state state state state state state state state state state state state state state state state state state state state state state state state state	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was into the operation procedure was into the operation procedure was into the operation process of mechanical resolutions of the operation of the Carl State State State State State Concentrations of the operation of the Electrical atoms that Electrical atoms that Electrical atoms that Badden atoms that a state balance at Badden atoms that Badden atoms that a state balance at Badden atoms that a state balance at Carlians induced function Contain state per togging Gane counter (g) Balance atoms (g) Balance atoms (g) Balance atoms (g) Balance atoms (g)	sediment s sediment s pOII plus ils s deta si was uses ned chemica zincon gr Riachuele x x x x x x x x x x x x x	amples wer CPP-MS am ails in Jarvi d for trace s Palma Sol XXX XX XX XXX XXX XXX XXX XXX XXX XXX	EC-ASS coupled with Thermo Xi mass spectromary, followed by its acribed by Solari et al. (2010). AVX5 acribed by Solari et al. (2010). AVX5 acribed by Solari et al. (2010). AVX5 acribed by Solari et al. (2010). AVX5 by mermalizing them with "SSL that retried on the zion gain subtack in the Mark Placement Advins, station, collision Salation, collision, forth temport Advins, station, collision, forth temport, stem Salation, scalaris, main, format pane, glack Advins, station, collision, forth temport, stem Salation, scalaris, main, format pane, glack Advins, station, collision, scalaris, forther pane, glack Advins, station, collision, scalaris, forther pane, glack Advins, station, collision, scalaris, forther pane, glack Advins, station, scalaris, forther pane, glack Advins, station, scalaris, forther pane, scalaris, forther dargenergy Advins, station, scalaris, forther dargenergy, scalaris, scalaris, forther dargenergy, scalaris, scalaris, forther dargenergy, scalaris, scalaris, forther dargenergy, scalaris, forther dargenergy, scalaris, scalaris, scalaris, forther dargenergy, scalaris, scalaris, scalaris, forther dargenergy, scalaris, scalaris, scalaris, sca	II Series quadrup Te 610 glass stand ment concentration of the stand of the concentration churcles and Palma intercord al zone hore subaquenous, surf al eas water (shalone fluid ter	and de- and de- and de- ans ans ans ans ans ans ans ans ans ans ans		Microtextu	Bit and state attention Bit attentint	Ziron g Richaldo X 22 22 23 24 24 24 24 25 25 26 26 27 27 26 27 27 27 27 27 27 27 27 27 27 27 27 27	Image: second	Ablacementerium Andran substrain, en Andran substrain, en Andran substrain, en Andran substrain, en Andran substrain, en Figherenge collisi Andran, substrain, fi regio-errage collisi Andran, substrain, fi regio-errage collisi Andran, substrain, fi regio-errage collisi Andran, substrain, fi regio-errage collisi Andran, substrain, fi Regio-errat, collisi Collision (alignetic, i en Substrain, substrain, fi Andran, Station, fi Andran, Station, fi Andran, Station, fi Andran, Station, fi Andran, Station, fi Andran, Station, fi Andrane, fi Andrane	er state state state state state state state state state state state state state state state state state state state state state state state state state	
	element concentration of 30 builts determined by a VC Elemental P the operation procedure was in (ICR). The United States Good ERC2 (Basals, Columbia Barer) Table 1 Montestered on the data and the state of the state of the state of the state Horizontal Columnia State of States Horizontal Columna States Horizontal P Horizontal Columna States Horizontal Columna States Horizontal Columna States Horizontal Columna States Horizontal Columna States Horizontal Columna States Horizontal Columna States Columna Stat	sediment s sediment s pOII plus ils s deta si was uses ned chemica zincon gr Riachuele x x x x x x x x x x x x x	amples wer CP–MS and in Jarvivey y Standard so Palma Sol XXX XX XX XX XX XX XX XX XX	EC-NAS coupled with Thermo Xi mass spectromary, followed by a crited by Solari et al. (2011). AVX5 a crited on the zone gain sufficient in the Parameter of the solar gain sufficient in the Parameter of the solar gain sufficient in the Autors, satistics, collision Salation, solarion, short semport following and the solarity of the solar solarity and the solarity of the solarity of the following collision, solarity, theory client following collision, solarity, theory client followi	II Series quadrup Te 610 glass stand ment concentration of the stand of the concentration churcles and Palma intercord al zone hore subaquenous, surf al eas water (shalone fluid ter	and de- and de- and de- ans ans ans ans ans ans ans ans ans ans ans			Advancement Advanceme	Ziron g Richaldo X 22 22 23 24 24 24 24 25 25 26 26 27 27 26 27 27 27 27 27 27 27 27 27 27 27 27 27	an	Addense Mathing on Addense Mathing on Address Mathi	er designed and a set of the set of th	

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

	referer	ice_id		
sample			▼	
sample_id	int		refere	nce
reference_id	int		reference	_id int
sample_id_in_paper	text		🔳 title	text
profile_name	text		authors	text
Iatitude	text		🔳 journal	text
Iongitude	text		📑 doi	text
age	text		💷 issn	text
depth	text		🔳 year	text
bioevent	text			

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.

AL.

17

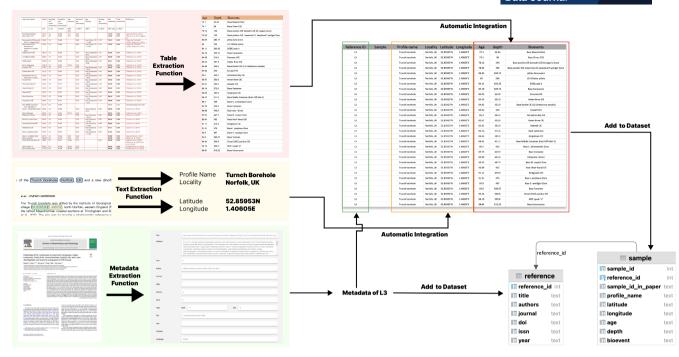


FIGURE 18 The example case workflow of using GeoDeepShovel.

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.

RMetS

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.

19

- 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: Datadriven 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 dataprocessing 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/10.1145/24636 76.2463680

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: <u>https://doi.org/10.1002/gdj3.186</u>