Accelerating Geoscience Research: An Advanced Platform for Efficient Multimodal Data Integration from Geoscience Literature

Zhixin Guo[1], Jianping Zhou[1], Guanjie Zheng[2], Xinbing Wang[1], and Chenghu Zhou[3]

[1]Department of Electronic Engineering, Shanghai Jiao Tong University (SJTU), [2]John Hopcroft Center for Computer Science, SJTU, [3]Institute of Geographic Sciences and Natural Resources Research, CAS

EGU General Assembly 2024

Abstract

In the era of big data science, geoscience has experienced a significant paradigm shift, moving towards a data-driven approach to scientific discovery. This shift, however, presents a considerable challenge due to the dispersed nature of data. To address this issue, we introduce a comprehensive, publicly accessible platform designed to facilitate the extraction and integration of multimodal data from geoscience literature, encompassing text, visual, and tabular formats. Our platform has been effectively applied in processing diverse domains, including mountain disaster data, global orogenic belt isotope data, and environmental pollutant data. This has facilitated substantial academic research, evidenced by developing **knowledge graphs based on debris flow data**, establishing a **global Sm-Nd isotope database**, and meticulous detection and analysis of environmental pollutants. As a core component of the Deep-Time Digital Earth (DDE) program, our platform has significantly contributed to the field, supporting forty geoscience research teams in their endeavors and processing over 40,000 documents. This accomplishment underscores the platform's capacity for handling large-scale data and its pivotal role in advancing geoscience research in the age of big data.

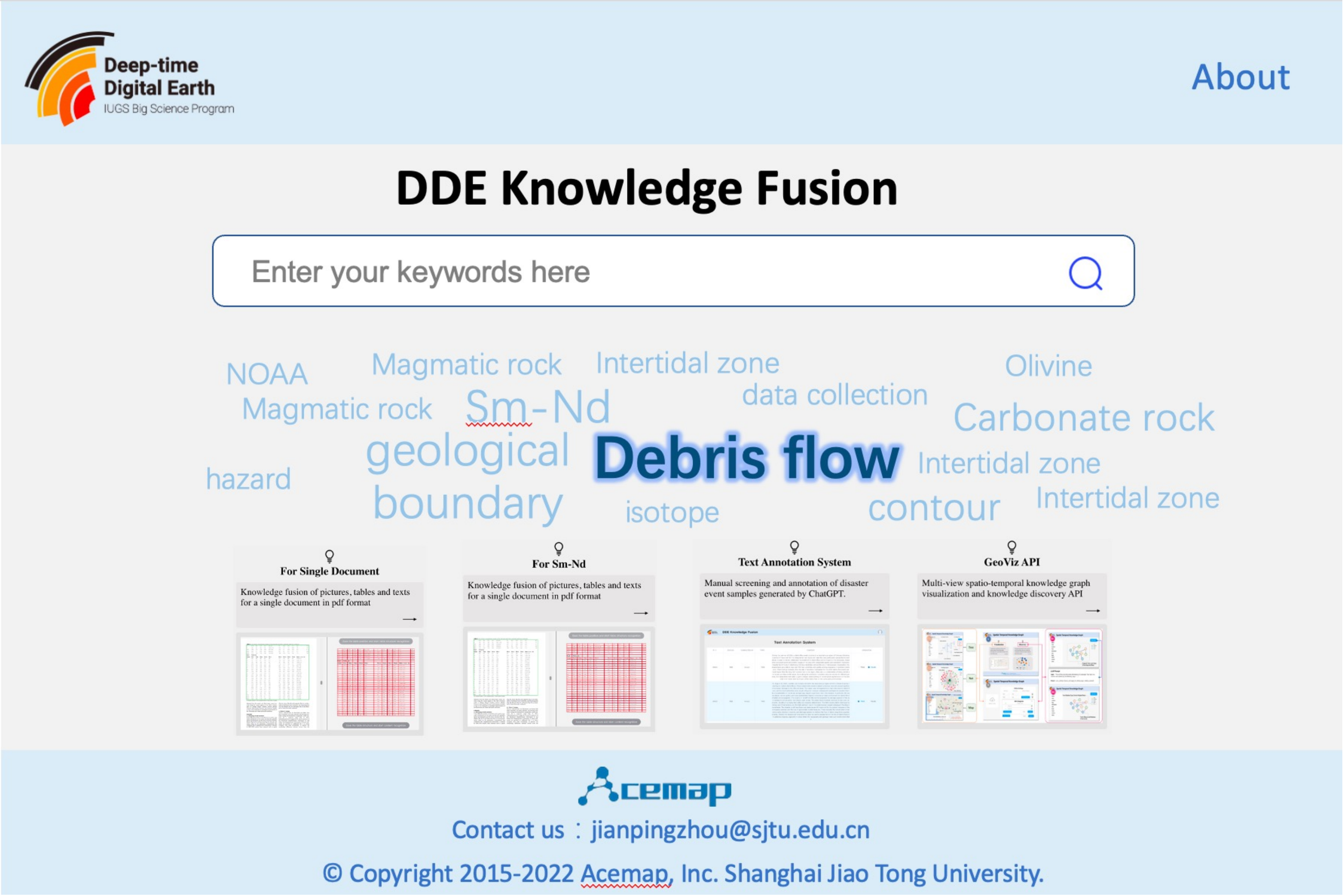
(@ Jianping Zhou)　(@ GeoKnowledgeFusion)

Background & Motivation

- Geological data exists in a multimodal form in a large number of academic literature, including tables, images, text and other forms.
- Geoscientific research in the era of big data often needs to mine knowledge from large amounts of geoscientific data and synthesize the correlations between them.
- However, it is very difficult to extract and normalize the heterogeneous data from multiple sources.
- Therefore, we hope to design a one-stop standardized process platform from literature to geoscientific knowledge to accelerate the research of geoscientists.

Our method & System

- a comprehensive, publicly accessible platform designed to facilitate the extraction and integration of multimodal data from geoscience literature.



DDE Knowledge Fusion

Enter your keywords here

For Single Document
Knowledge fusion of pictures, tables and texts for a single document in pdf format

For Sm-Nd
Knowledge fusion of pictures, tables and texts for a single document in pdf format

Text Annotation System
Manual screening and annotation of disaster event samples generated by ChatGPT.

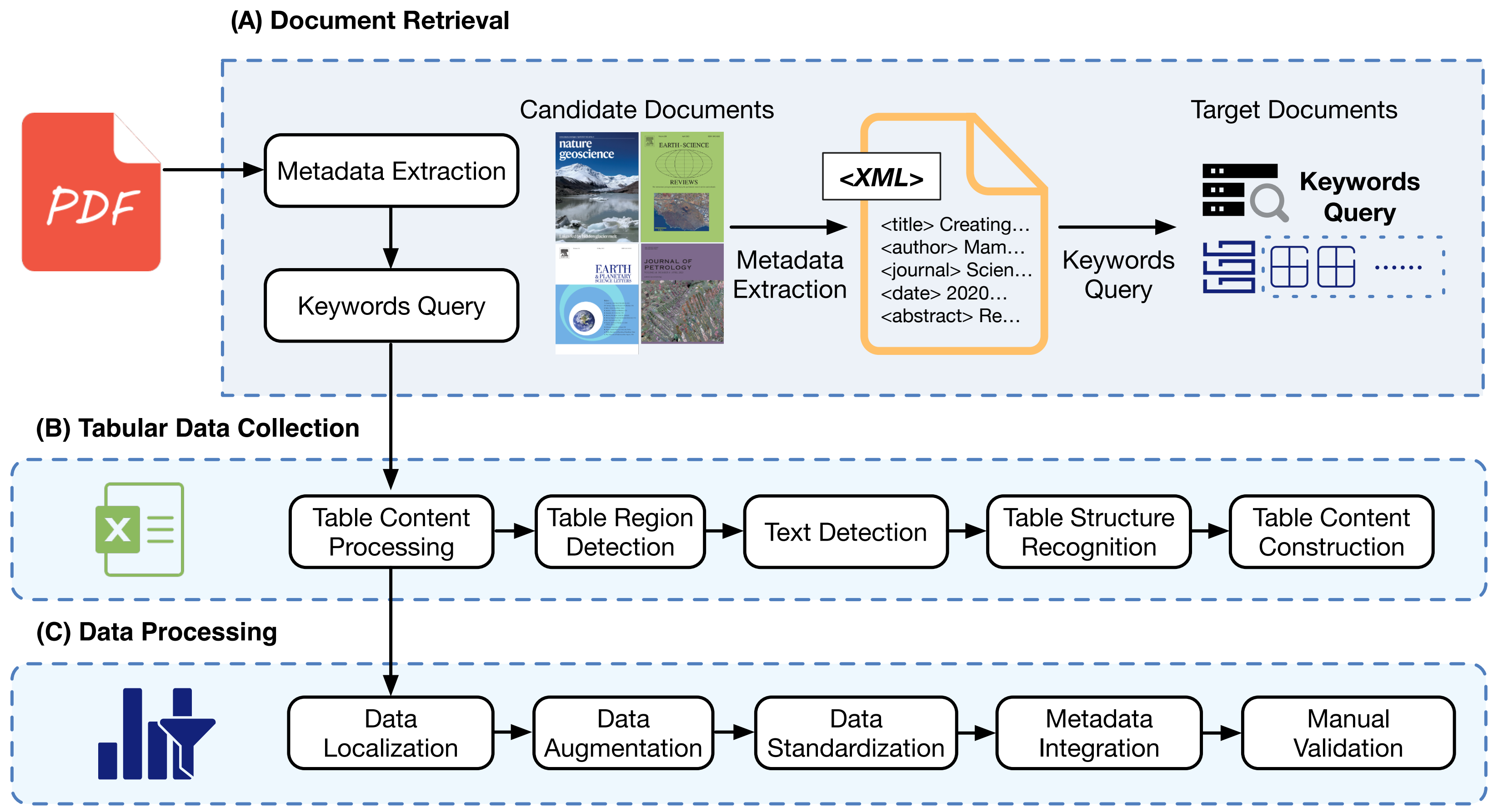
GeoViz API
Multi-view spatio-temporal knowledge graph visualization and knowledge discovery API

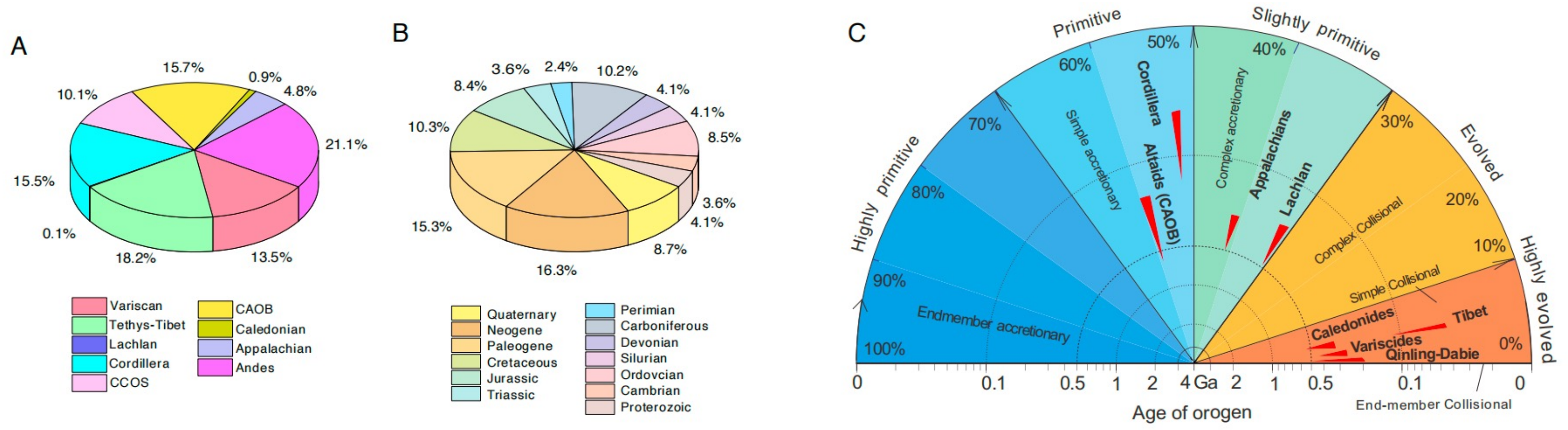
Contact us：jianpingzhou@sjtu.edu.cn

© Copyright 2015-2022 Acemap, Inc. Shanghai Jiao Tong University.

Case 1: The Global Sm-Nd Isotope Tabular Database

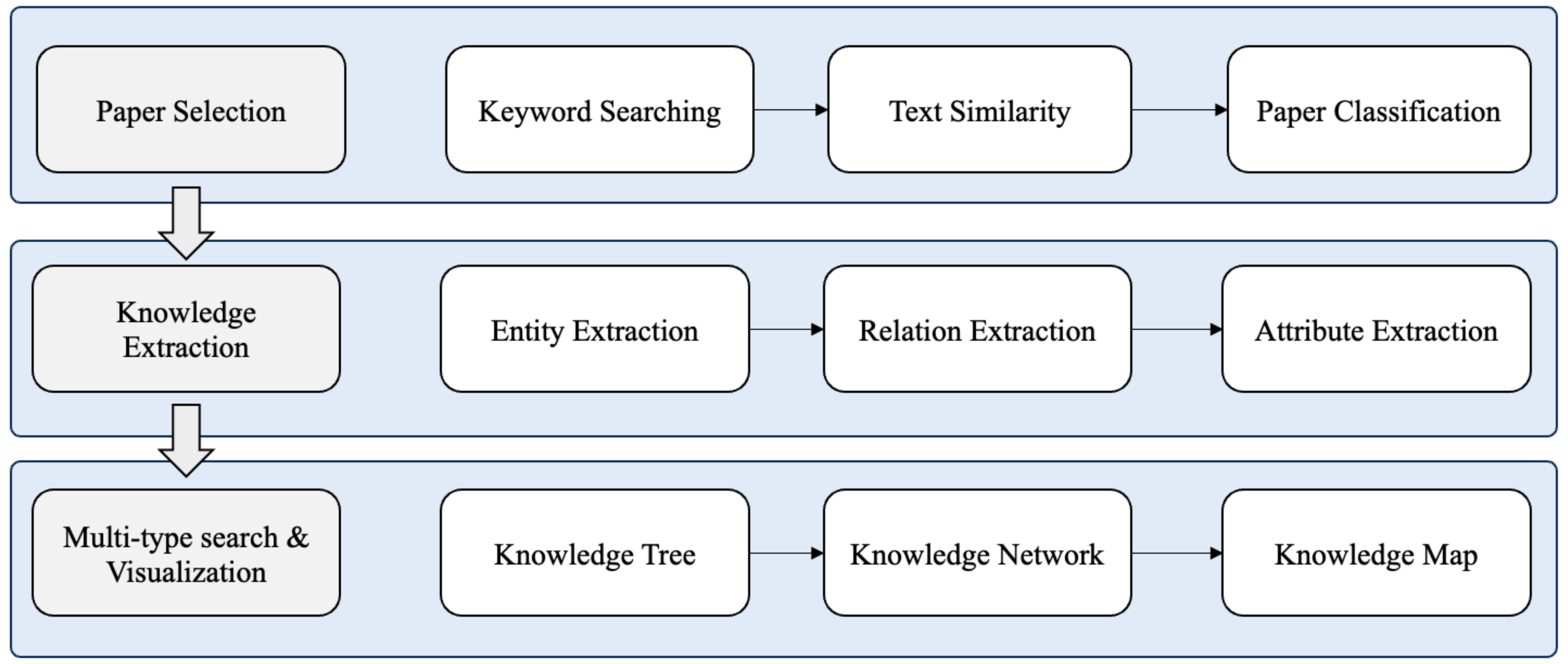
- Sm-Nd isotope data are widely available in geoscientific literature tables, necessitating the efficient integration of this data into a database through automated extraction and fusion method.
- We collected 270,527 academic papers from 33 types of geology journals, screened 10,548 Sm-Nd isotope data, parsed 42,432 tables, screened 4,807 valid tables, and finally parsed 17,334 valid data.

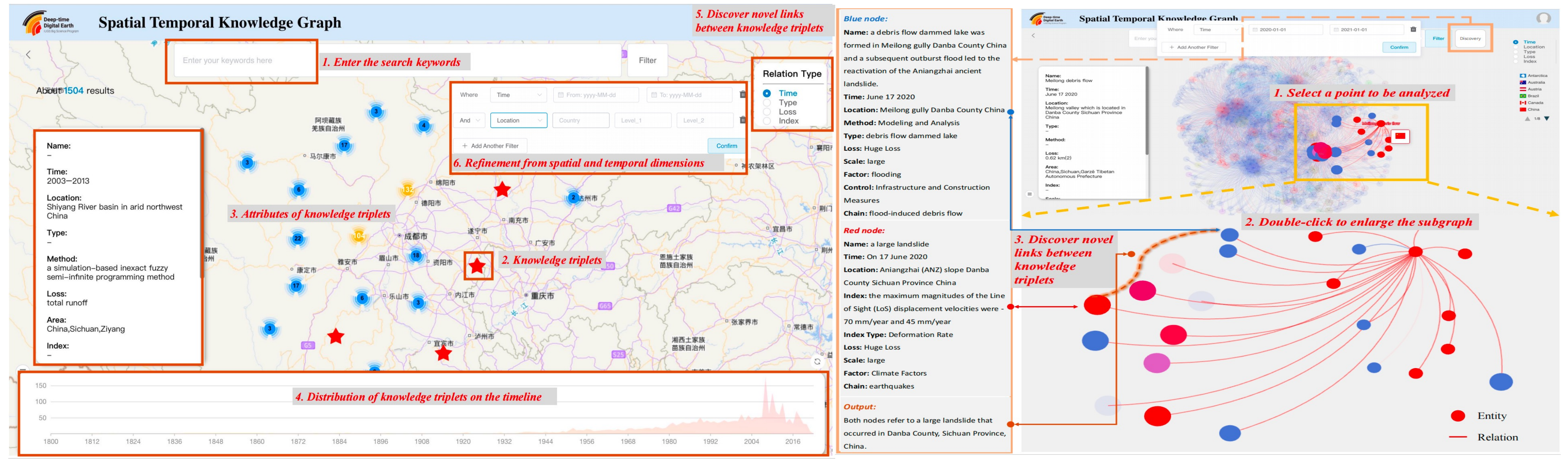




Case 2: Debris Flow Knowledge Graph

- The global debris flow lacks systematic and refined information on the details of the hazard, and the mechanism that causes the hazard, which are widely available in the texts of geoscience literature.
- We screened 55,360 articles of debris flow literature from 136 types of mountain disaster journals, and extracted 18 types of debris flow hazard indicators.





Blue node:
Name: a debris flow dammed lake was formed in Meilong gully Danba County China and a subsequent outburst flood led to the reactivation of the Aniangzhai ancient landslide.
Time: June 17 2020
Location: Meilong gully Danba County China
Method: Modeling and Analysis
Type: debris flow dammed lake
Loss: Huge Loss
Scale: large
Factor: flooding
Control: Infrastructure and Construction Measures
Chain: flood-induced debris flow

Red node:
Name: a large landslide
Time: On 17 June 2020
Location: Aniangzhai (ANZ) slope Danba County Sichuan Province China
Index: the maximum magnitudes of the Line of Sight (LoS) displacement velocities were - 70 mm/year and 45 mm/year
Index Type: Deformation Rate
Loss: Huge Loss
Scale: large
Factor: Climate Factors
Chain: earthquakes

Output:
Both nodes refer to a large landslide that occurred in Danba County, Sichuan Province, China.

Name:
–
Time:
2003–2013
Location:
Shiyang River basin in arid northwest China
Type:
–
Method:
a simulation–based inexact fuzzy semi–infinite programming method
Loss:
total runoff
Area:
China,Sichuan,Ziyang
Index:
–

Spatial Temporal Knowledge Graph
Enter your keywords here
1. Enter the search keywords
Filter
About 1504 results
5. Discover novel links between knowledge triplets
Relation Type
Time
Type
Loss
Index
Where
Time
And
Location
Add Another Filter
Confirm
6. Refinement from spatial and temporal dimensions
3. Attributes of knowledge triplets
2. Knowledge triplets
4. Distribution of knowledge triplets on the timeline
150
100
50
1800
1812
1824
1836
1848
1860
1872
1884
1896
1908
1920
1932
1944
1956
1968
1980
1992
2004
2016
1. Select a point to be analyzed
2. Double-click to enlarge the subgraph
3. Discover novel links between knowledge triplets
Entity
Relation

Related Paper

[1] Zhixin Guo, Jianping Zhou, et al., "Towards Controlled Table-to-Text Generation with Scientific Reasoning", ICASSP, 2024.

[2] Zhixin Guo, et al. "Sm-Nd Isotope Data Compilation from Geoscientific Literature Using an Automated Tabular Extraction Method", 2024.

Session: ITS1.8/TS9.1– Navigating the Digital Earth: Advancements in Earth Systems and Geophysical Inference(EGU24-13715)