GAKG

https://gakg.acemap.info/

Multimodal Geoscience Academic Knowledge Graph

Github: https://github.com/davendw49/gakg

GAKG: A Multimodal Geoscience Academic Knowledge Graph

Cheng Deng[1], Yuting Jia[1], Chong Zhang[1], Jingyao Tang[1], Hui Xu[1], Luoyi Fu[1], Weinan Zhang[1], Haisong Zhang[2], Xinbing Wang[1], Chenghu Zhou[3]

[1]Shanghai Jiao Tong University, [2]Tencent AI Lab

[3]Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences

{davendw,hnxxjyt,yiluofu}@sjtu.edu.cn,hansonzhang@tencent.com,zhouch@lreis.ac.cn

Presented by Cheng Deng

http://www.big-cheng.com

https://github.com/davendw49

davendw@sjtu.edu.cn

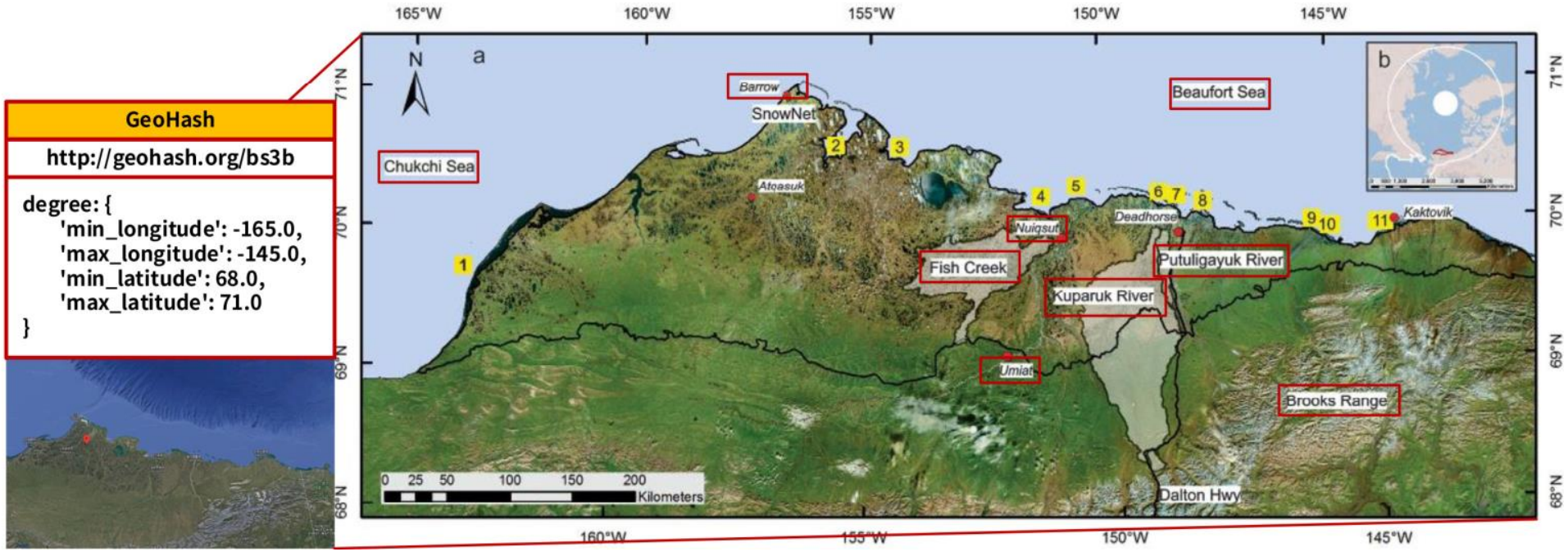
Deep-time Digital Earth
IUGS Big Science Program

cemap

SHANGHAI JIAO TONG UNIVERSITY

CIKM 2021

30th ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT



A large quantity of literatures

Papers, a major mean to disseminate knowledge

geoscience papers contain abundant multimodal data

Papers thus have time and spatial characteristics

DBpedia WIKIDATA Freebase yago select knowledge Google ProBase

Multi-domain Knowledge graph

Microsoft Academic Graph

Established: June 5, 2015

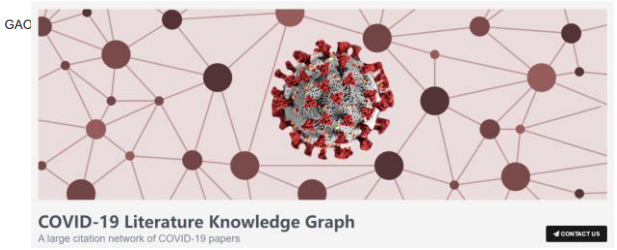
cemap

AMiner

Academic Knowledge graph

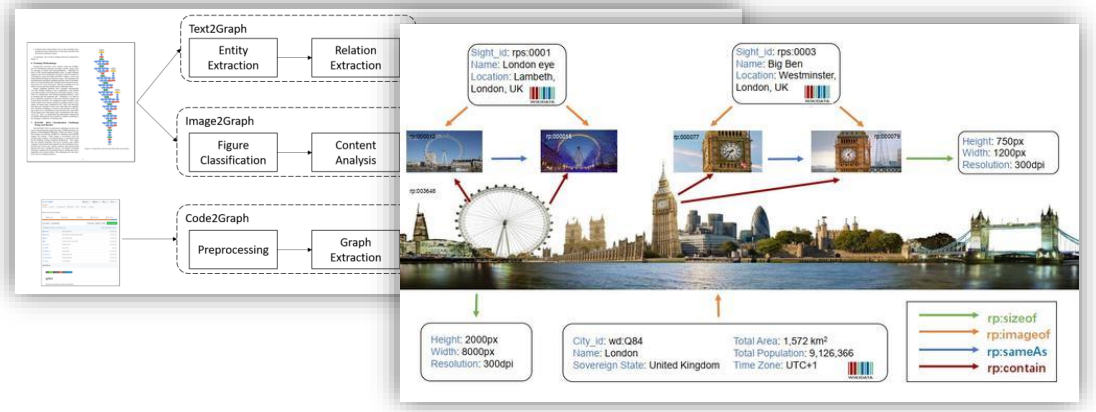
A common sense geographic knowledge base for GIR

ZHANG Yi†, GAO



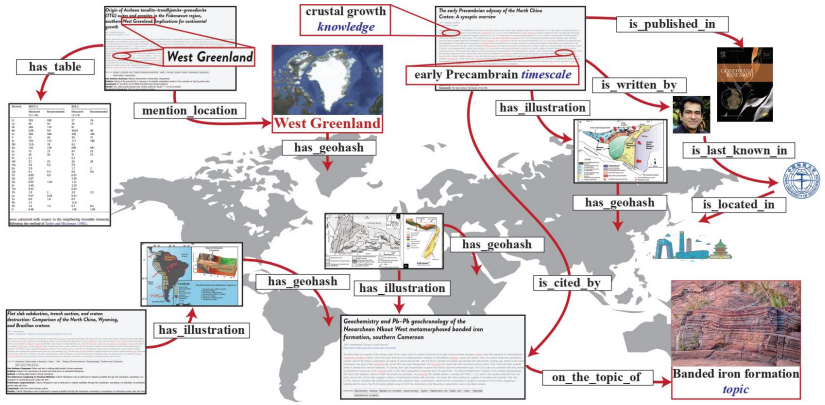
GeoNames

Domain Knowledge graph



Multimodal Knowledge graph

GAKG



Multimodal Geoscience Academic Knowledge Graph

So, Link the knowledge together!

crustal growth *knowledge*

Origin of Archean tonalite–trondhjemite–granodiorite (TTG) suites and granites in the Fiskenæsset region, southern West Greenland: Implications for continental growth

West Greenland

has_table

mention_location

West Greenland

has_geohash

The early Precambrian odyssey of the North China Craton: A synoptic overview

early Precambrain *timescale*

is_published_in

is_written_by

has_illustration

is_last_known_in

is_located_in

has_geohash

has_geohash

has_geohash

has_illustration

is_cited_by

has_illustration

Flat slab subduction, trench suction, and craton destruction: Comparison of the North China, Wyoming, and Brazilian cratons

Geochemistry and Pb–Pb geochronology of the Neoarchean Nkout West metamorphosed banded iron formation, southern Cameroon

on_the_topic_of

Banded iron formation *topic*

GAKG's schema-graph consists of **11** concepts connected by **19** relations. Five of them (*has_concluded, has_designed, is_located_in, has_developed and earn_in_the_way_of*) have a upper class relation *acer:mention_knowledge*. Since **GAKG** is the union of academic concepts and their relations, we manage GAKG as linked open data (LOD), we provide *#sameAs* axioms linking to the entities in other datasets, **271 thousands** in total.

The Graph base namespace (Graph IRI) is ***https://www.acekg.cn***, all the concepts and relations shared.

To our knowledge, GAKG is currently the **largest** and **most comprehensive** geoscience academic knowledge graph, consisting more than **68 million triples, including 8 millions concepts and 41 million links.**

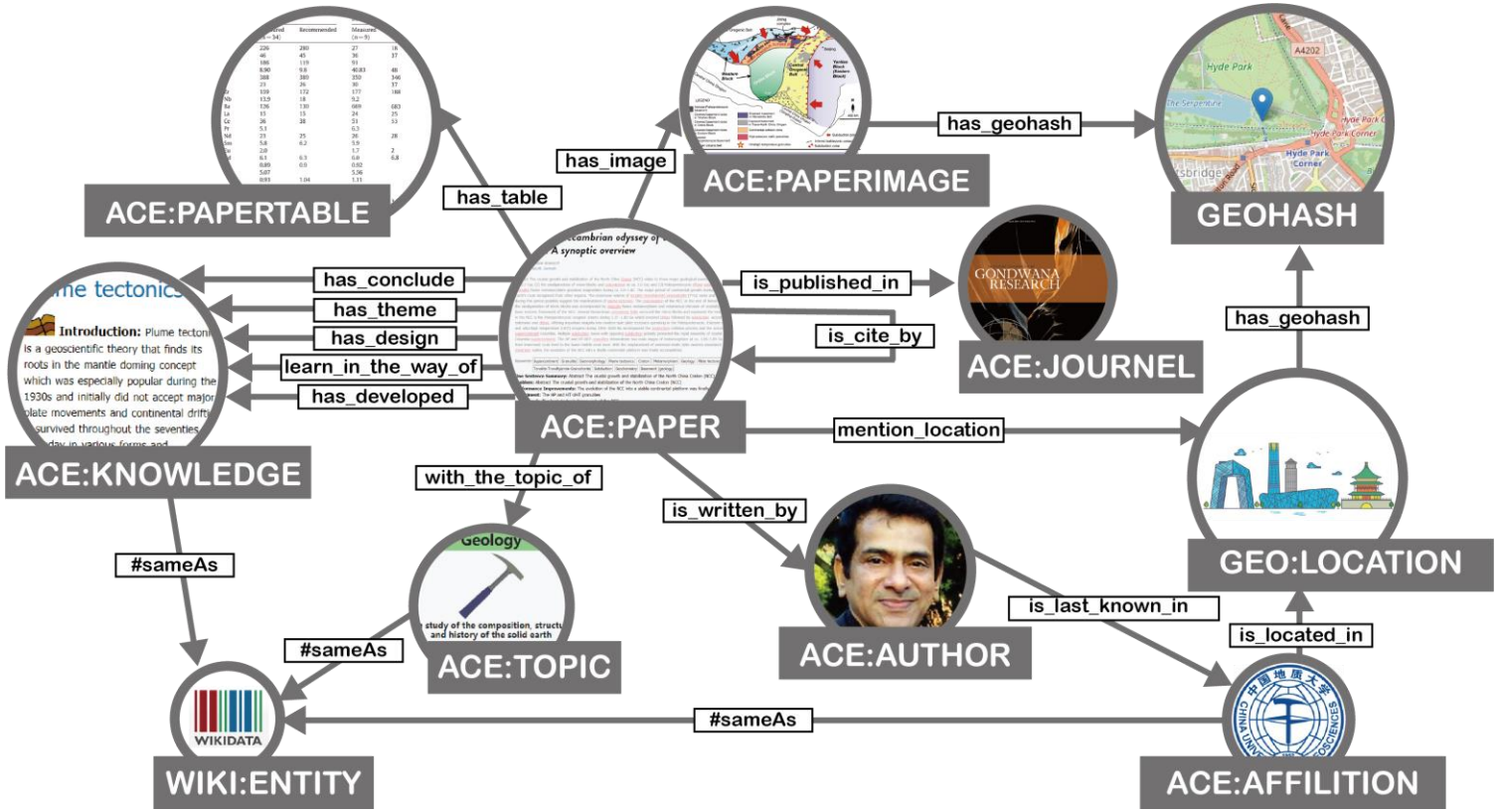
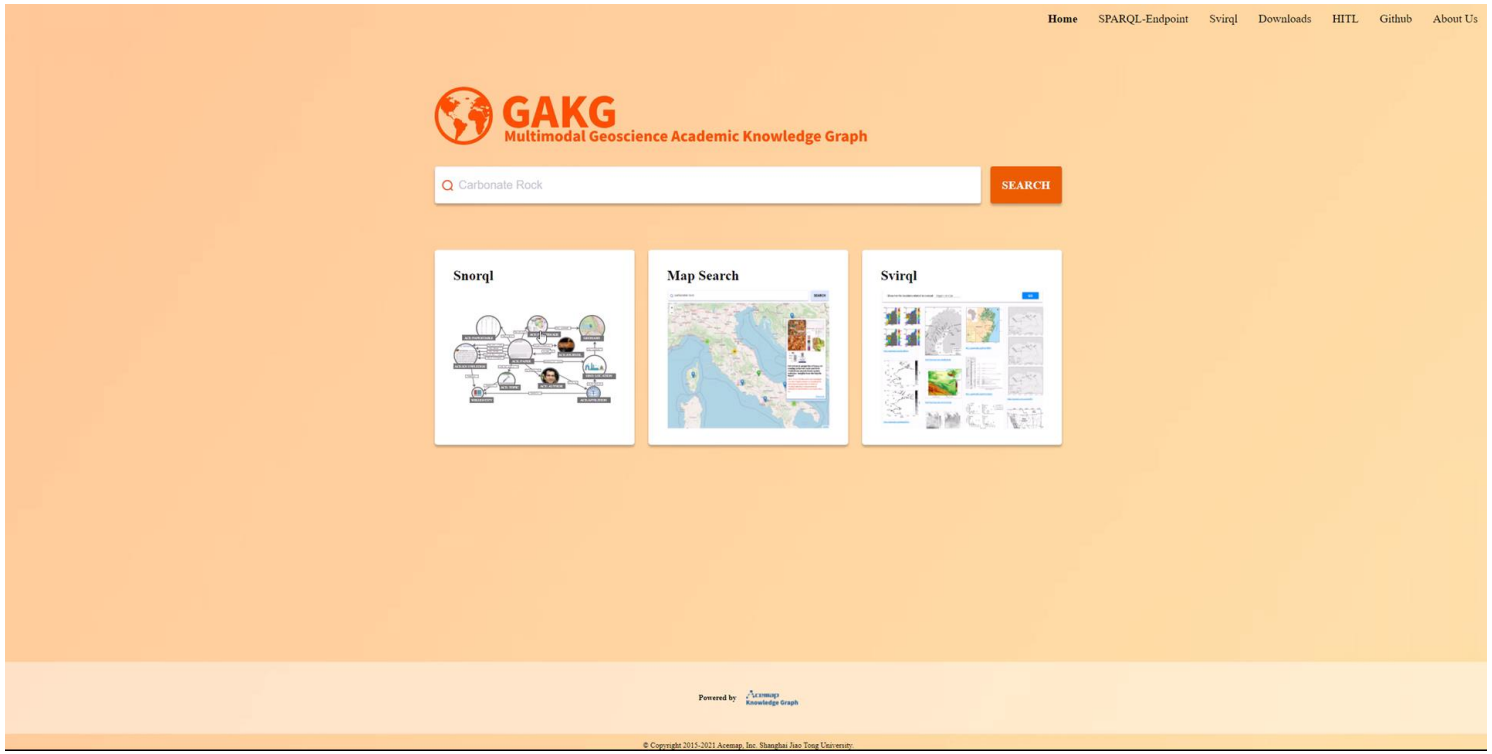


Table 1: Statistics of GAKG Concepts (Up to May 30, 2021).

Concept	Count	Concept	Count
paper	1,122,094	knowledge	62,576
author	908,933	illustration	3,562,816
affiliation	27,175	papertable	760,054
topic	765,184	location	784,279
journal	194	geohash	996,731
timescale	1,701	**Total**	8,991,737

Table 2: Statistics of GAKG Relations (Up to May 30, 2021).

Relation	Count	Relation	Count
is_cited_by	17,704,495	mention_knowledge	704,899
on_the_topic_of	10,401,972	mention_location	759,260
is_written_by	3,547,077	has_geohash	1,021,870
is_published_in	1,122,094	mention_timescale	1,120,398
is_last_known_in	662,850	in_the_period_of	189
is_located_in	25,019	before	155
has_illustration	3,562,816	#sameAs	271,156
has_table	760,054	**Total**	41,664,304



GAKG SPARQL: https://www.acekg.cn/sparql

GAKG Snorql: https://snorql.acemap.cn/

- Ref: SNORQL: https://github.com/kurtjx/SNORQL
- Ref: Virtuoso: https://virtuoso.openlinksw.com/

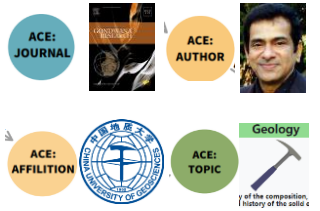
GAKG Core Instance Level

DDE Scholar

PDF Parsing, Information Extraction, Explicit Semantic Analysis

Deep Shovel

academic metadata



Abstract

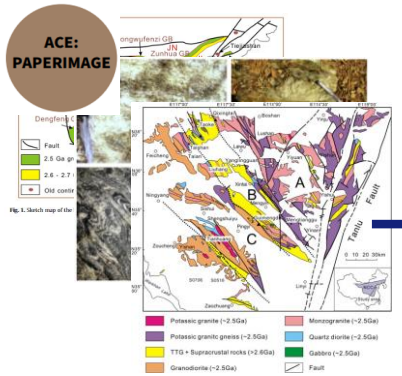
The crustal growth and stabilization of the North China Craton (NCC) relate to three major geological events in the Precambrian: (1) a major phase of continental growth at ca. 2.7 Ga; (2) the amalgamation of micro-blocks and cratonization at ca. 2.5 Ga;

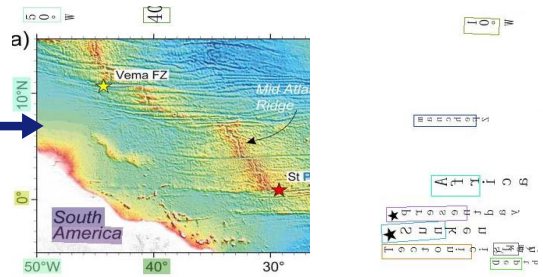
…

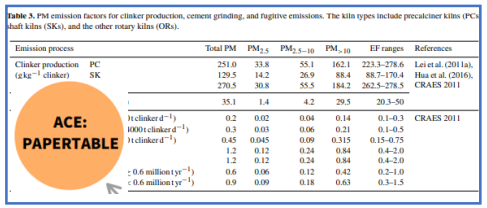
1. Introduction

…

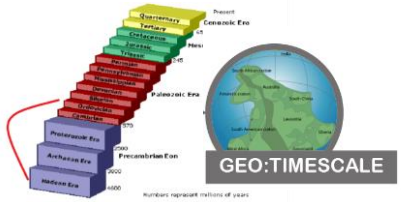
However, the mechanism of aggregation of the Early Precambrian continental blocks to form a supercraton is equivocal, and the various models proposed take into consideration rifting, density-inversion, back arc basin formation, arc–arc accretion, subduction of oceanic lithosphere





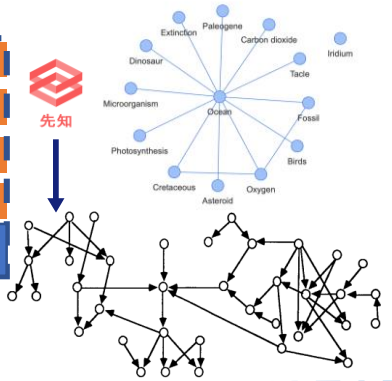


time scale data

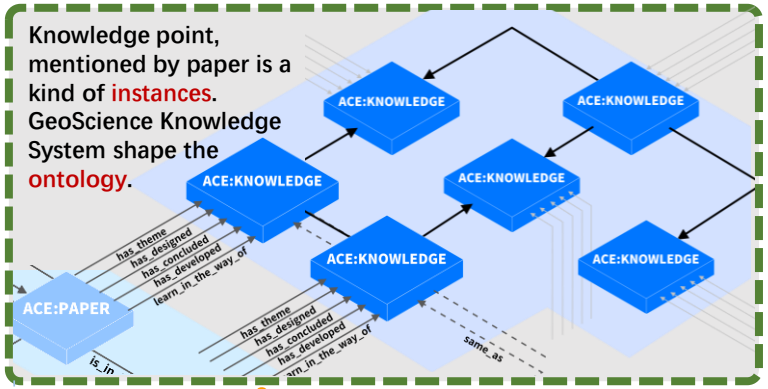


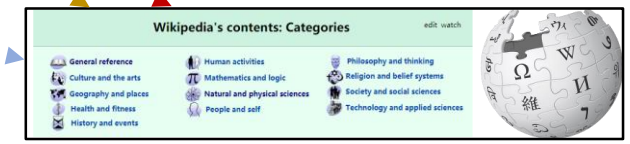
GAKG Core Ontology Level

crustal growth
stabilization
basin formation
arc-arc accretion
Entity co-occurency



Knowledge point, mentioned by paper is a kind of instances. GeoScience Knowledge System shape the ontology.

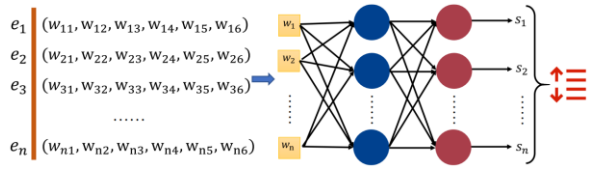




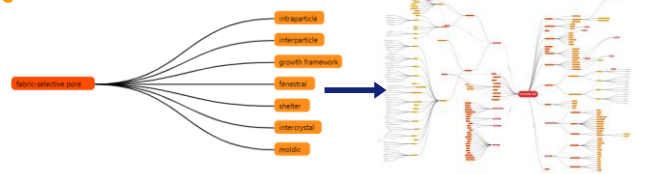
WIKI: ENTITY

GeoScientist in the Loop

ACE: KNOWLEDGE



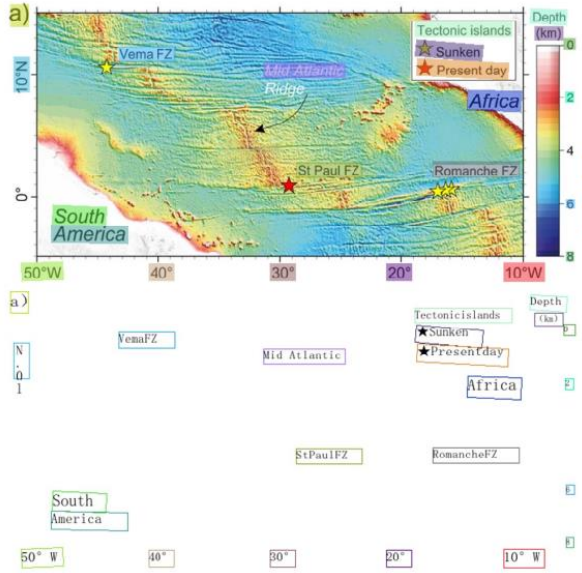
Find the key phrase most related to the text.



GAKG on the Map

West Greenland LOC

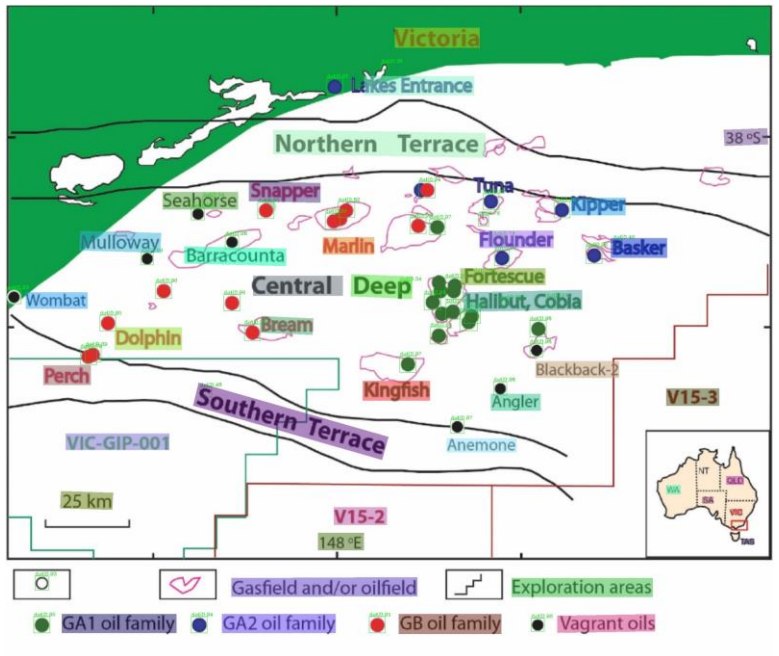
has_geohash

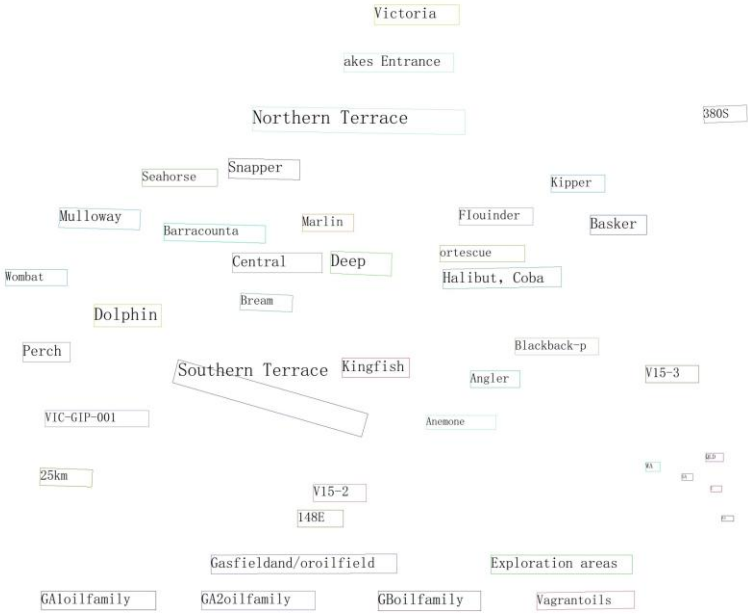


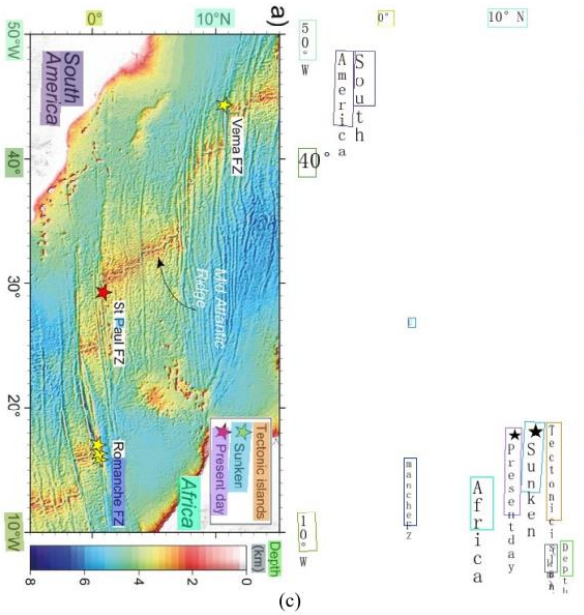
(a)

Text	Score	X	Y
a)	0.945	16	8
Depth	1.000	657	12
Tectonicislands	0.999	515	29
(km)	0.998	662	35
★Sunken	0.997	517	50
0	0.982	699	49
VemaFZ	0.970	149	59
N.0l	0.759	20	72
Mid Atlantic	0.910	328	79
★Presentday	1.000	518	73
Africa	0.999	580	112
2	1.000	701	115
StPaulFZ	0.994	368	202
RomancheFZ	0.975	537	201
South	1.000	68	254
6	0.999	702	247
America	0.990	66	279
8	0.999	702	312
50°W	0.950	29	326
40°	0.999	186	327
30°	1.000	336	327
20°	0.987	480	327
10°W	0.971	624	326

(b)







(c)

Text	Score	X	Y
10°N	0.900	291	29
0°	0.997	131	30
、e	0.524	349	21
50°W	0.920	37	41
America	1.000	85	77
South	0.981	108	77
40°	1.000	37	196
30°	1.000	37	345
StPau	0.997	160	378
Present day	0.994	285	537
Sunken	0.998	308	538
Tectonicislands	1.000	331	531
RomancheFZ	0.996	159	556
Africa	0.999	249	592
10°W	0.980	36	635
（km）	0.993	328	672
Depth	1.000	348	668
C	0.645	50	712

(d)

An Example of Illustrations Coordinates Extraction.

(a) is the result of OCR,
(b) highlights the recognition error,
(c) is the result of OCR after using a rule-based method adjusting image,
and (d) highlights the corrected coordinates.

AceMap DeepScholar

Question Sets

Machine Reading

BERT based Neural Machine Reading

output

question	origin_res
aim	[1] An integrated study of topography, bathymetry, high-resolution aeromagnetic data, and structural observations
problem	significant Cenozoic fault activity in SE Ireland.
method	The integration of offshore high resolution aeromagnetic data and structural interpretation of the Kish Bank Basin
methor_base	The integration of offshore high resolution aeromagnetic data and structural interpretation of the Kish Bank Basin
result	onshore Ireland has also been affected by significant tectonic activity and exhumation during the Cenozoic.

ESA, entity extraction

elasticsearch

dbpedia+acemap

dbpedia+acemap

Q (Relation)	Entity(Object)
aim	Integrator (2247163)
	Integral(2375312)
	Bathymetry(2336508)
problem	Cenozoic(2312161)
	Activant(2032866)
	Activate(1110643)
method	Kish Bank(848923)
result	Tectonics(2299601)
	Ireland(60999)
	Affection(2075186)

wikidata

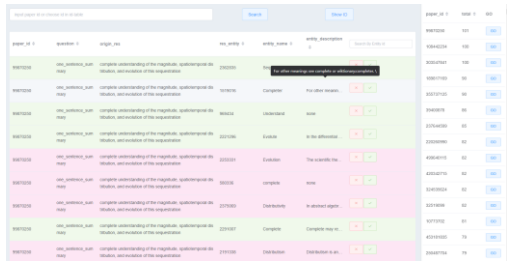
Q (Relation)	Entity(Object)
aim	Topography(Q134435)
	Observation(Q94575845)
	Bathymetry(Q810826)
problem	Ireland(Q22890)
	Fault(P5095)
method	Kish Bank(Q52556663)
	Basin(Q29382870)
result	Ireland(Q22890)
	Exhumation(Q1898510)

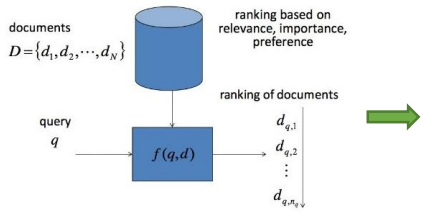
Entity Features:
1. Word length
2. Num. of the Word
3. Sim. with title
4. Sim. with question
5. Sim. with answer
6. Complexity of the Word

Human Annotation

LTR-LambdaRank

Build: mention_knowledge





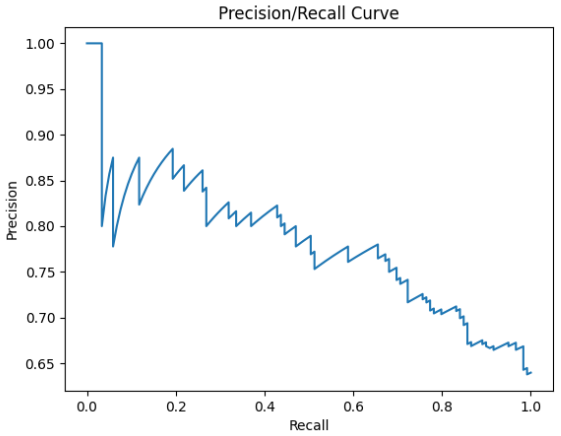
Virtuoso

neo4j

D2RQ

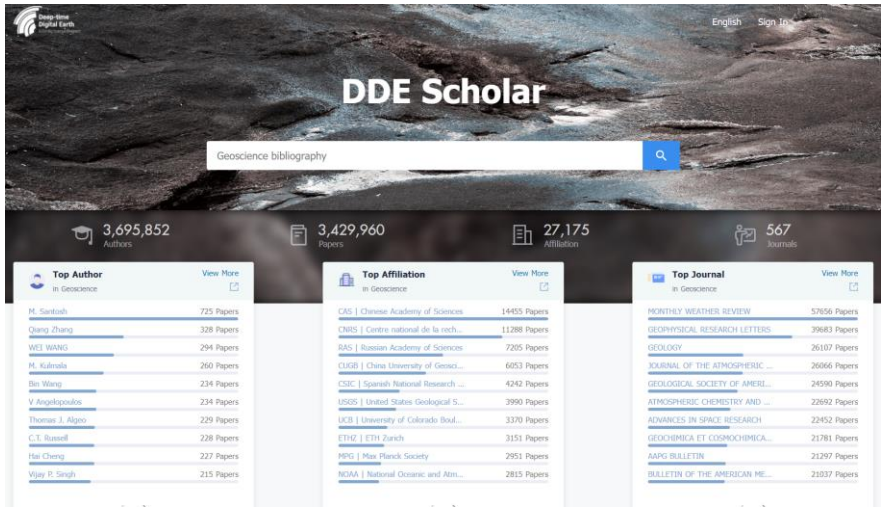
Linking the Ontology and Instances

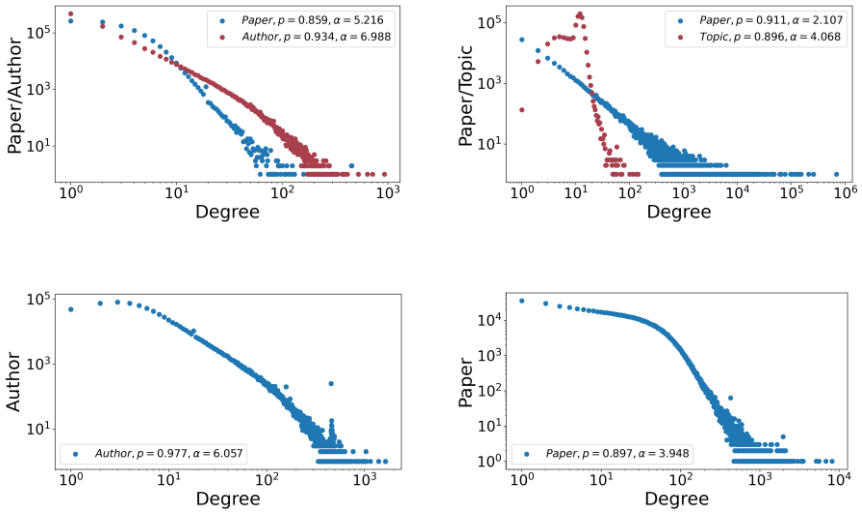
Entities List

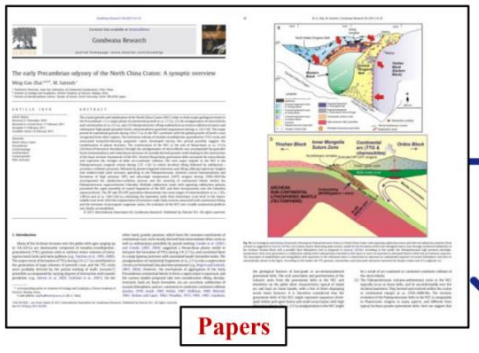


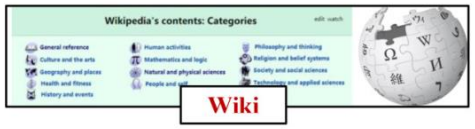
Models	precision
ESA without Supervision	0.63
ESA + Learn2rank (after 1st loop HITL)	0.79 (↑25%)

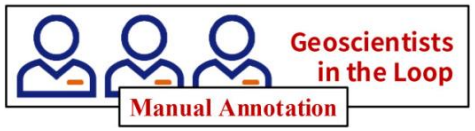
Models	precision
Autophrase without Supervision	0.59
Autophrase + Learn2rank (after 1st loop HITL)	0.78 (↑32%)

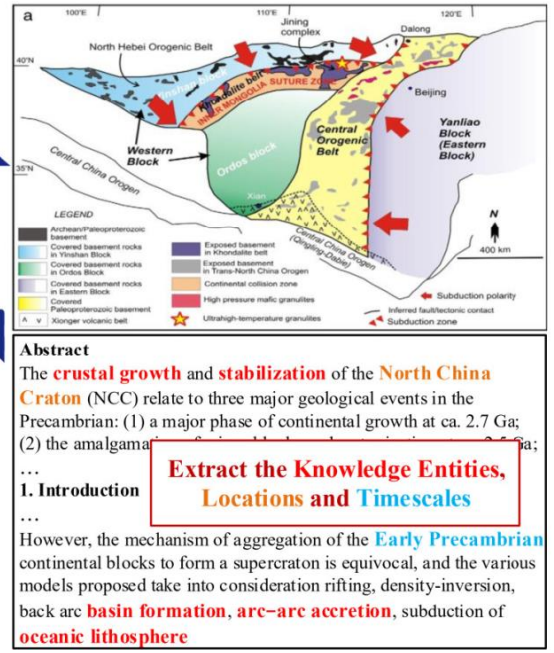


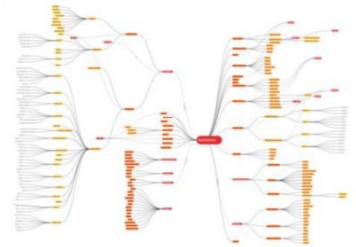


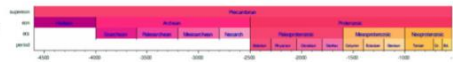












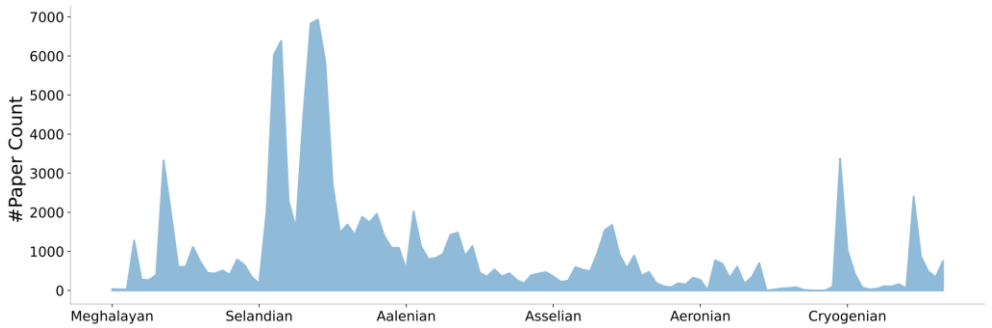


Figure 8: Papers Distribution along with Geologic Era.

Benchmark-CD

Table 5: Statistics of Community Detection Benchmarks.

Benchmarks	Number of communities	Nodes	Edges	Nodes in Largest WCC	Edges in Largest WCC	Nodes in Largest SCC	Edges in Largest SCC	Average Cluster Coefficient	Triangles	Diameter
GPCN	194	842,121	16,034,510	838,219 (0.995)	16,031,892 (0.999)	0	0	0.0699	38,789,469	176
GACN	194	860,280	5,381,861	752,718 (0.875)	5,282,032 (0.972)	752,718 (0.875)	5,282,032 (0.972)	0.6897	43,502,542	15
Email-Eu-core	42	1,005	25,571	986 (0.981)	25,552 (0.999)	803 (0.799)	24,729 (0.967)	0.3994	105,461	7
CORA	7	2,708	5,429	2,485 (0.918)	2,604 (0.493)	13 (0.005)	14 (0.003)	0.1314	1,630	15
DBLP (Collaboration Network)	2,547	317,080	1,049,866	317,080 (1.000)	1,049,866 (1.000)	317,080 (1.000)	1,049,866 (1.000)	0.6324	2,224,385	21
Amazon (Product Network)	5,000	334,863	925,872	334,863 (1.000)	925,872 (1.000)	334,863 (1.000)	925,872 (1.000)	0.3967	667,129	44

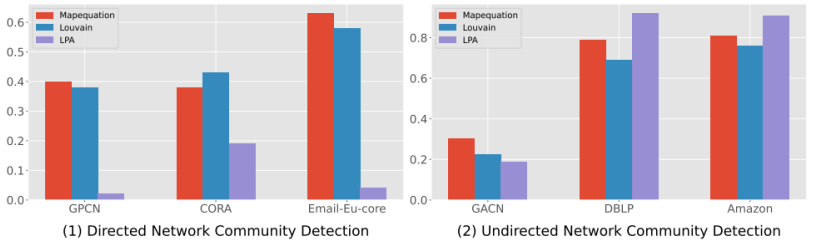


Figure 9: Community Detection Evaluation Results.

Benchmark-KE

Table 6: Statistics of KRL Benchmarks.

Benchmark	relation	entity	triple
FB15K	1,345	14,951	483,142
WN18	18	40,943	141,442
GA16K	10	16,363	151,662

Table 7: Results of Link Prediction Task.

Models	FB15K		WN18		**GA16K**	
	MR	hit@10	MR	hit@10	MR	hit@10
RESCAL	683	0.441	1,163	0.528	4,300	0.001
TransE	125	0.471	251	0.892	280	0.320
TransH	84	0.585	303	0.867	337	0.325
RotatE	40	0.884	309	0.959	214	0.366
SimpIE	74	0.876	412	0.947	311	0.260

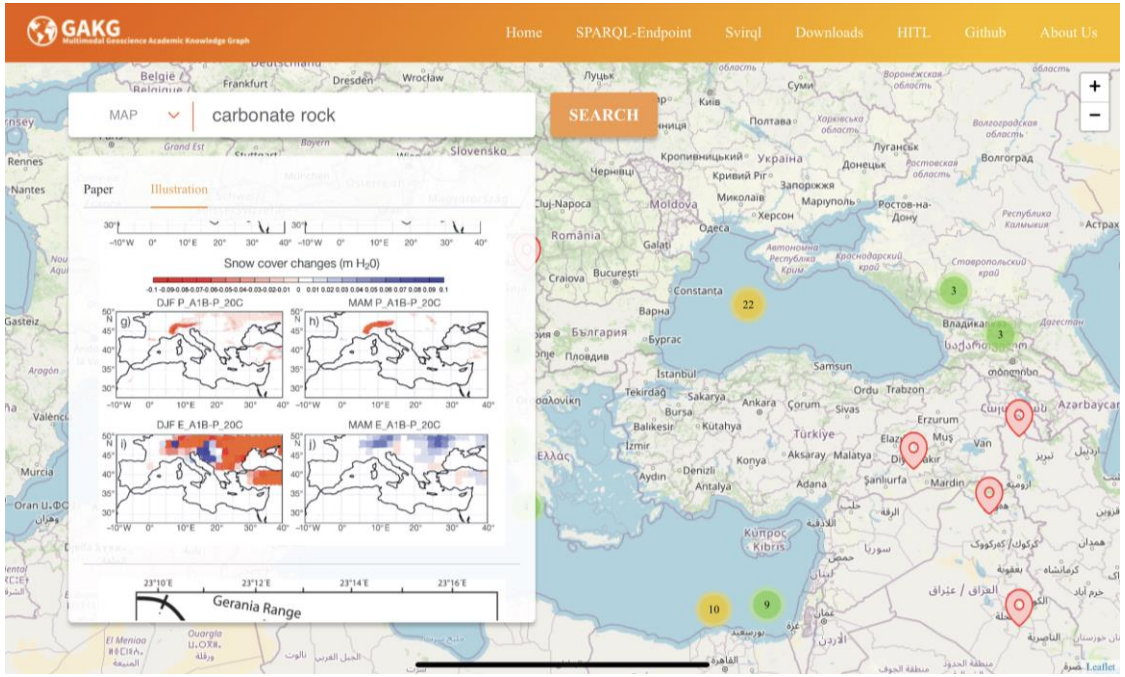
Contribution

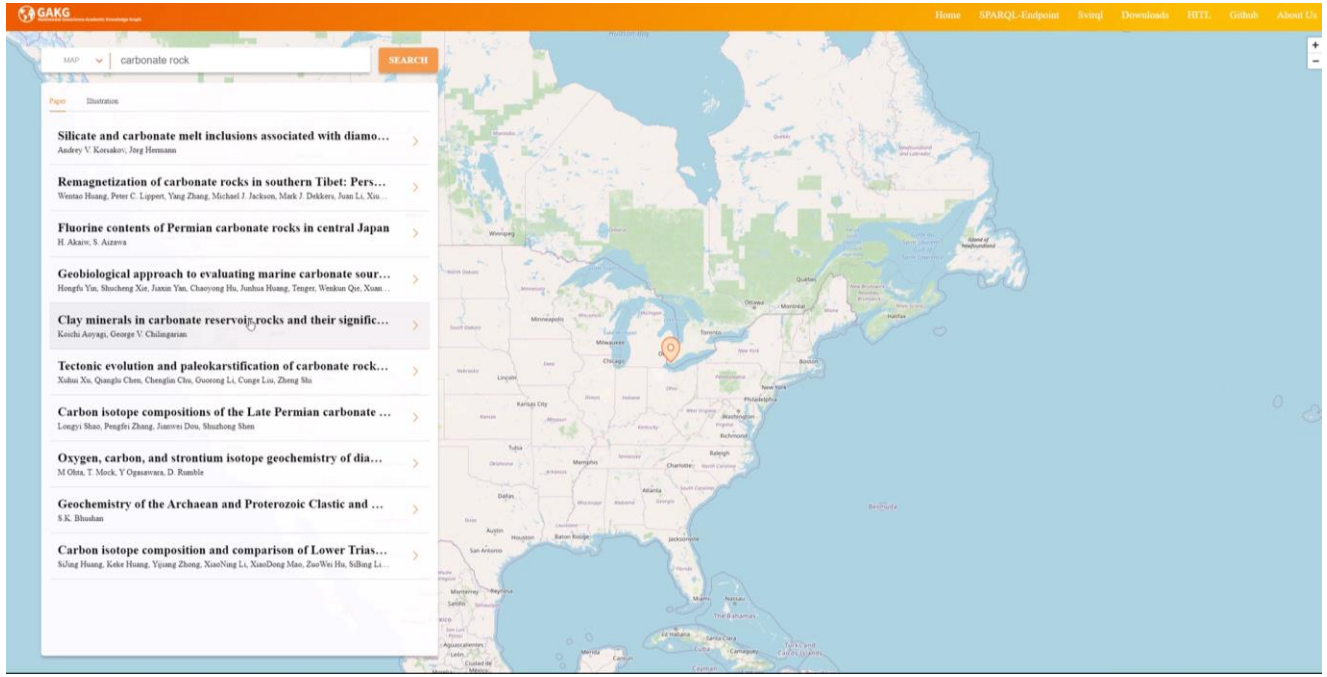
- We propose a multimodal **GeoScience Academic Knowledge Graph (GAKG)** framework by fusing papers' images, text, and bibliometric data.
- With a world map, all the illustrations, text, and geologic time scale extracted from the selected geoscience papers can be highly connected to the geographical information
- We put forward a Human-In-the-Loop knowledge extraction pipeline to extract paper's knowledge entities and mapping them to a crowd-sourcing knowledge taxonomy.

Observation

- *Availability.*
  - *Dump files*
  - *SPARQL Endpoint*
  - *Snorql Query System*
  - *Papers on the Map*
  - *Source Code on Github*
- *Quality.*
  - *Data sources: Acemap, DBpedia, DDE*
  - *Verified partially by Geoscientists*
- *Limitations*
  - *Lack of annotation during the Human-In-The-Loop*

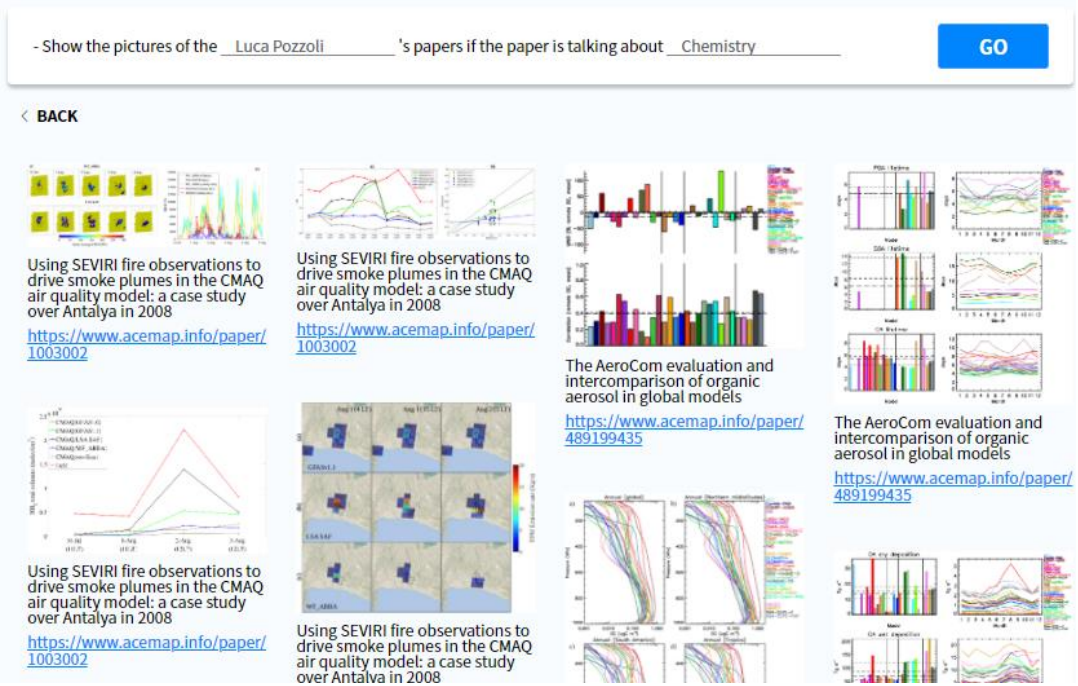
- **Geographic Information Retrieval.** We provide a knowledge based search engine on a geographical map for the literature of geoscience.
- If the researcher drags the window, the distribution of the papers will change accordingly.
- **LINK: https://gakg.acemap.info/**

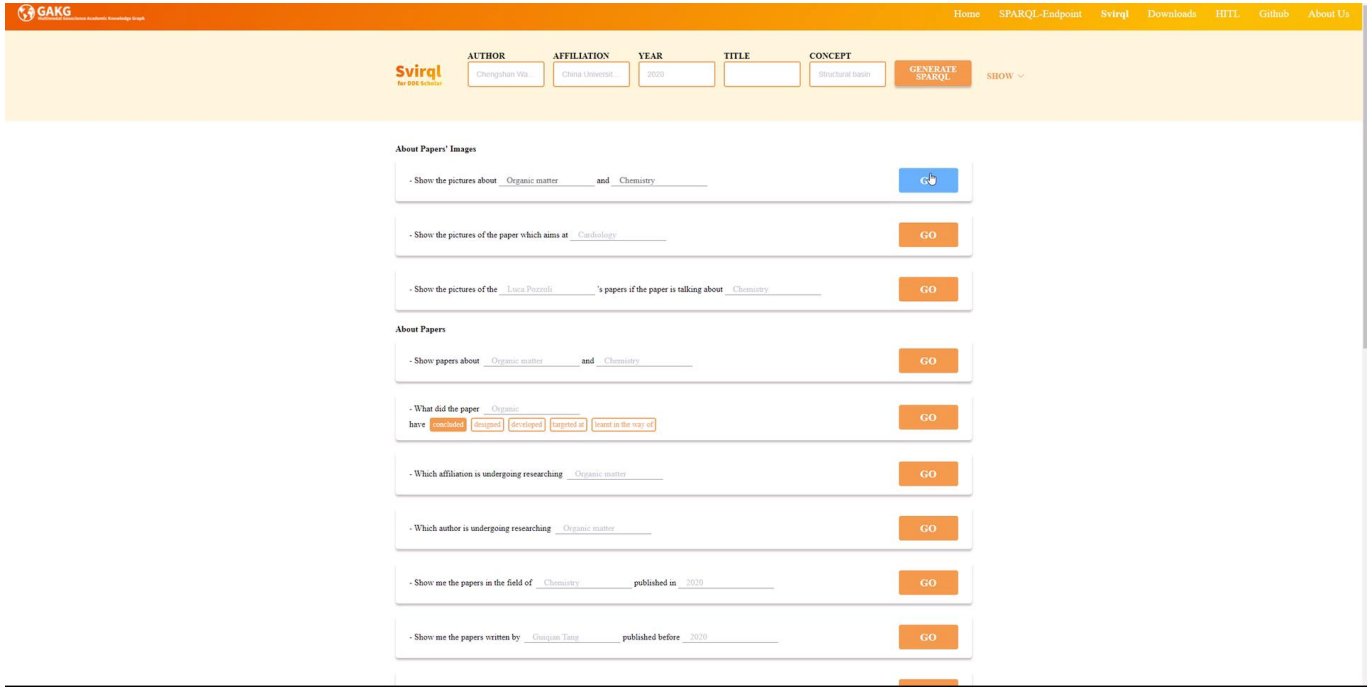




- ***Geoscience KBQA.*** Based on GAKG, research can know more information about the relation between papers. These template-based queries can be applied in scientific research and academic communication. These questions are also generally inextricable by existing Q&A systems and search engines.

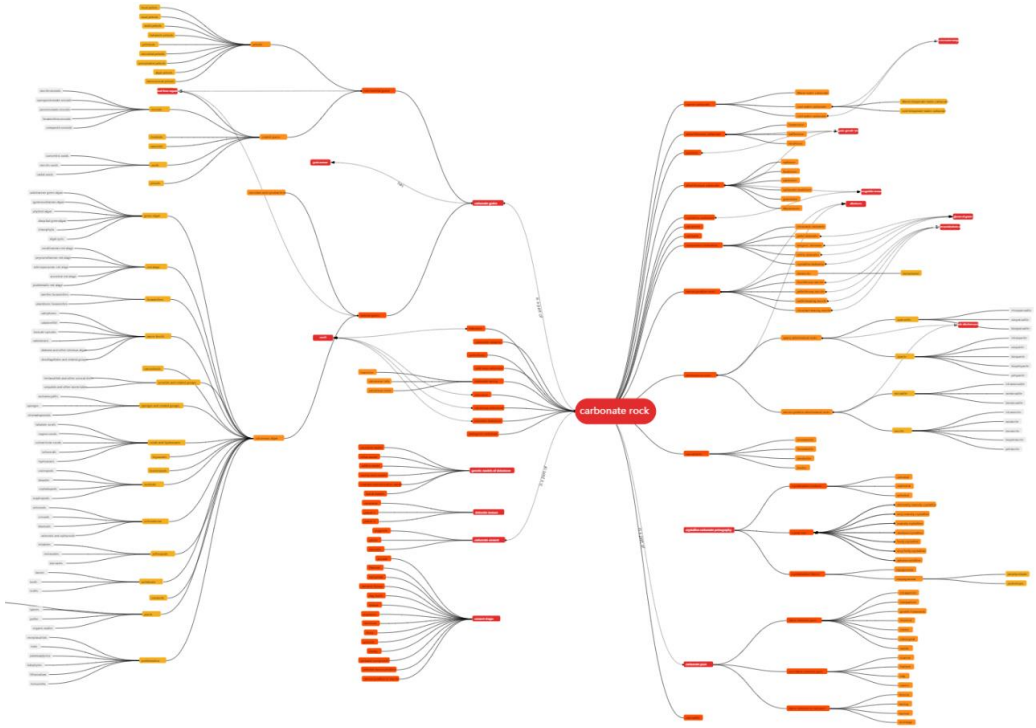
- One-hop queries, such as returning papers targeting a particular topic,
- Two-hop queries, such as querying illustrations in a specific field,
- Three-hop queries, such as querying geographic locations that a certain affiliation often studies,
- Four-hop queries, such as querying the relationship between geographic locations and affiliations.





Future Works

- *Academic Knowledge System Construction with Human-In-the-Loop.*
- *Social Community Detection in Geoscience.*
- *Scientific Articles' Geographical Information Extraction.*



GAKG

https://gakg.acemap.info/

Multimodal Geoscience Academic Knowledge Graph

Github: https://github.com/davendw49/gakg

Thank You!

GAKG: A Multimodal Geoscience Academic Knowledge Graph

Cheng Deng[1], Yuting Jia[1], Chong Zhang[1], Jingyao Tang[1], Hui Xu[1], Luoyi Fu[1], Weinan Zhang[1], Haisong Zhang[2], Xinbing Wang[1], Chenghu Zhou[3]

[1]Shanghai Jiao Tong University, [2]Tencent AI Lab

[3]Institute of Geographical Science and Natural Resources Research, Chinese Academy of Sciences

{davendw,hnxxjyt,yiluofu}@sjtu.edu.cn,hansonzhang@tencent.com,zhouch@lreis.ac.cn

Presented by Cheng Deng

http://www.big-cheng.com

https://github.com/davendw49

davendw@sjtu.edu.cn

Deep-time Digital Earth
IUGS Big Science Program

cemap

上海交通大學
SHANGHAI JIAO TONG UNIVERSITY

CIKM 2021

30th ACM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT