# Biologically-oriented mud volcano database: muddy_db

**Alexei Remizovschi** [1] , **Rahela Carpa** Corresp. [2]

[1] Department of Molecular Biology and Biotechnology, Faculty of Biology and Geology, Babes-Bolyai University, Cluj-Napoca, - Select -, Romania
[2] Department of Molecular Biology and Biotechnology, Faculty of Biology and Geology, Babes Bolyai University, Cluj-Napoca, - Select -, Romania

Corresponding Author: Rahela Carpa
Email address: rahela.carpa@ubbcluj.ro

Mud volcanoes (MVs) are naturally occurring hydrocarbon hotbeds with continuous methane discharge, contributing to global warming. They host microbial communities adapted to hydrocarbon oxidation. Given their research value, MVs still represent a niche topic in microbiology and are neglected by hydrocarbon-oriented research. All the data regarding MVs is sporadic and decentralized. To mitigate this problem, we built a custom Natural Language Processing pipeline (muddy_mine), and collected all the available MV data from open-access articles. Based on this data, we built the muddy_db database. The muddy_db represents the first biologically oriented database rendered as a user-friendly web app. This database includes all the relevant MV data, ranging from microbial taxonomy to hydrocarbon occurrence and geology. The muddy_mine and muddy_db tools are licensed under the GPLv3.

muddy_db R Shiny web app: https://muddy-db.shinyapps.io/muddy_db/

muddy_db R package: https://github.com/TracyRage/muddy_db

muddy_mine Conda package: https://github.com/TracyRage/muddy_mine

# Biologically-oriented mud volcano database: muddy_db

**Remizovschi Alexei**[1] **and Rahela Carpa**[1]

[1]**Department of Molecular Biology and Biotechnology, Babeș-Bolyai University, Cluj-Napoca, Cluj, Romania**

Corresponding author:
Rahela Carpa[1]

Email address: rahela.carpa@ubbcluj.ro

## ABSTRACT

Mud volcanoes (MVs) are naturally occurring hydrocarbon hotbeds with continuous methane discharge, contributing to global warming. They host microbial communities adapted to hydrocarbon oxidation. Given their research value, MVs still represent a niche topic in microbiology and are neglected by hydrocarbon-oriented research. All the data regarding MVs is sporadic and decentralized. To mitigate this problem, we built a custom Natural Language Processing pipeline (muddy_mine), and collected all the available MV data from open-access articles. Based on this data, we built the muddy_db database. The muddy_db represents the first biologically oriented database rendered as a user-friendly web app. This database includes all the relevant MV data, ranging from microbial taxonomy to hydrocarbon occurrence and geology. The muddy_mine and muddy_db tools are licensed under the GPLv3.
muddy_db R Shiny web app: https://muddy-db.shinyapps.io/muddy_db/
muddy_db R package: https://github.com/TracyRage/muddy_db
muddy_mine Conda package: https://github.com/TracyRage/muddy_mine

## INTRODUCTION

Mud volcanoes (MVs) represent hydrocarbon discharging landforms (Mazzini and Etiope, 2017). They are distributed worldwide in both marine and terrestrial environments (Milkov, 2000). The most distinctive feature of MVs is recurrent methane emission. Due to methane emissions, MVs contribute extensively to global warming (Etiope et al., 2009).

MV genesis is mainly caused by a naturally mediated process - kerogen maturation (Vandenbroucke and Largeau, 2007). Therefore, the surrounding area of MVs can provide valuable data regarding both aerobic and anaerobic hydrocarbon microbial oxidation (Cheng et al., 2012).

Over the years, MVs research has been mainly focusing on anaerobic oxidation of methane (AOM) and the implicit interaction between sulfate-reducing bacteria and methane oxidizing archaea (ANME) (Bose et al., 2013; Cui et al., 2014). In addition to AOM research, MVs were also investigated in the context of hydrocarbon research. A myriad of MV studies discussed the thermogenic and biogenic origin of the methane (Etiope et al., 2009; Sano et al., 2017).

Despite these studies, the biological aspect of MVs is still a niche and unexplored topic. The biological data regarding MVs are sporadic and mostly biased towards AOM. Even worse, the already available data is not centralized. Due to the lack of the MV-dedicated database, researchers still analyze MV as a singular entity and avoid comparative studies between MVs with geographically distinct locations.

Meanwhile, mainstream biomedical fields have extensively employed natural language processing (NLP) techniques to mine meaningful data from ~~the~~ research articles (Wang et al., 2020). Simultaneously, the number of databases related to biomedical fields is considerable (Luo et al., 2016). Niche environmental science fields have not caught up. ~~Lacking the possibility to~~ mine environmental-oriented articles and build field-specific databases ~~delays~~ the publication of the meta-analyses or any comparative studies.

Fortunately, democratic NLP models and tools have been published over the last years. Some of them can be easily used by environmental scientists with limited computer science (CS) experience, for example, the spaCy library, ScispaCy models, and S2ORC database (Honnibal and Johnson, 2015; Neumann et al., 2019; Lo et al., 2020).

~~Cumulatively, the latest~~ advancements in NLP ~~can~~ provide opportunities for consolidating and promoting niche environmental topics such as MV microbiology.

~~Given these facts, we built~~ the first biologically oriented mud volcano database, muddy_db~~, a~~ niche database ~~that~~ consolidates all ~~the~~ relevant biological ~~data, which~~ will be of great use for researchers specializing in bacterial hydrocarbon oxidation or MV microbiology. ~~Collaterally, our custom~~ pipeline can serve as a methodological blueprint for ~~research collectives~~ interested in ~~NLP and building~~ their own specialized databases.

# METHODS

To collect all the available data regarding the biological aspects of MVs, we had to exclusively rely on open-access articles. ~~Having these articles, we could freely mine~~ all the biologically oriented tokens, including taxonomy-, chemicals-, geology-, and MV-specific terms. ~~Additionally, we built a custom mining pipeline - muddy_mine (Fig. 1).~~ The~~ scope of~~ muddy_mine ~~is to provide and enrich the~~ muddy_db database ~~with relevant MV data~~.

## Data collection

We used the S2ORC (20200705v1) database to collect open-access articles. S2ORC represents a centralized database that includes 12.7 million articles with a fully preserved paper structure. S2ORC is quite comprehensive and includes niche environment science articles (Lo et al., 2020). Given these facts, we extracted all the available MV-related titles (N=118 total, N=115 deduplicated) from the S2ORC.

## Token extraction and muddy_mine pipeline

Having MV articles, we proceeded with token extraction (i.e. the extraction of the terms of interest) using the muddy_mine pipeline.

Taxonomy extraction represented a difficult challenge due to the fact that we intended to collect as many tokens as possible. To overcome this problem, we used the spaCy library (2.3.2), third-party ScispaCy NLP models (0.3.0), and the most recent NCBI Taxonomy database (20 November, 2020) (Honnibal and Johnson, 2015; Neumann et al., 2019; Schoch et al., 2020). First, we extracted all the taxon tokens using en_core_sci_sm ScispaCy model (2.2.5). Second, we checked those tokens against a local NCBI Taxonomy database. Third, we counted the extracted tokens. ~~The higher the counting number,~~ the more likely the token was explicitly discussed in the article.

By interating this ~~third~~-step algorithm, we managed to centralize MV-specific taxonomy on all the possible levels: phylum, class, order, family, and genus.

~~The other non-taxonomy tokens were also extracted with the above-mentioned algorithm,~~ We extracted and counted the tokens related to the following categories: chemistry (inorganic ions, hydrocarbons), geology (geological periods, minerals), MV terminology (ANME, methanogenesis type), and experimental methods (PCR types, amplified genes, chromatography). The ~~comprehensive~~ list ~~of categories~~ can be ~~consulted by visiting~~ the muddy_db repository.

The raw output of the muddy_mine pipeline ~~represents~~ a set of csv tables with MV data.

## Building muddy_db database

~~By~~ obtaining muddy_mine raw output, we can advance to the next step - building a ~~user-friendly database~~. To ~~create this kind of database~~, we created a Shiny web app, entitled muddy_db. In order to build it, we used the following R packages: shiny (1.5.0), semantic.dashboard (0.2.0), and golem (0.2.1) (Filip and Igras, 2021; Chang et al., 2021; Fay et al., 2021). This app includes all the output generated by the muddy_mine pipeline. Specifically, it displays ~~the counted~~ tokens extracted both from ~~the integral~~ article bodies (N=57) and abstracts (N=115). Additionally, we added an annotated map, which displays the geographical distribution of MVs and their affiliated research metadata.

## System requirements

~~We would like to mention that~~ muddy_mine pipeline was designed to run on systems with modest memory requirements. We achieved this feature by using Python generators. Intel Core i3 (3rd Gen) 3217U / 1.8 GHz processor (Intel, USA) and 4GB RAM system was used to build muddy_db. The mining process ~~lasted~~ 24 h.

## RESULTS

The ~~scope~~ of the muddy_db is to gather all the available MV biologically relevant data and include it in a user-friendly database (Fig. 2). First, we collected all the known taxa associated with MVs. The muddy_db includes data regarding archaeal and bacterial taxonomy on all the possible taxonomy levels. This particularity can facilitate the detection of microbial consortia patterns. Second, we gather information regarding metabolic pathways, geology, hydrocarbon availability, and experimental methods performed on MV sediments (Fig. 3). This information can guide specialists to implement appropriate research strategies.

## DISCUSSION

MVs are considered to be one of the settings where ~~the~~ early life evolved (Pons et al., 2011). They sustain a plethora of bacterial metabolic pathways, ranging from methane oxidation and synthesis to sulphate reduction (Kleindienst et al., 2014; Cheng et al., 2012). These pathways and their affiliated microbial communities could provide valuable data regarding (1) origin of microbial life, (2) the effect of the naturally occurring methane discharging systems on global warming, (3) the contribution of microbial consortia to oil souring (Gieg et al., 2011; Etiope et al., 2009; Pons et al., 2011) ~~As we can see, the~~ accumulation of MVs data could enhance ~~our~~ knowledge regarding topics that range from fundamental studies to ecology and engineering.

Given these facts, MVs should be the main focus of hydrocarbon-oriented research and ecology. Unfortunately, data regarding the biological aspects of MVs are scarce. Additionally, the data already gathered are not combined in a dedicated database. The lack of a specialized MV database determines mud volcano microbiology to be a niche and neglected topic in environmental science.

Biomedical fields have always represented the cutting-edge subset of natural science, which actively implement CS techniques, and are tightly intertwined with the big data term (Luo et al., 2016). Simultaneously, the implementation of CS methods in niche environmental fields lags. To both apply CS methods in an environmental context and chronically mitigate the data deficient field of MV microbiology, we created a muddy_mine NLP pipeline and muddy_db database.

The creation of muddy_db tools aims to create a platform, that would provide sufficient data to perform meta-analyses or comparative studies of the MVs. Specifically, we hope that muddy_db would facilitate the discovery of atypical taxonomic patterns and point out the influence of geography on MVs characteristics. Simultaneously, muddy_mine represents a reproducible example of a mining technique applied in the context of environmental studies. As a result, muddy_mine could encourage researchers to mine their data of interest, being free from any field of study constraints.

In addition to positive sides of our projects, we would like to address the evident limitations of muddy_mine pipeline. We were limited to only to open-access articles which were found in the S2ORC. Currently, we are constricted to this corpus due to the following facts:

1. Journals use various article xml encoding standards such as TEI-XML and JATS-XML. Therefore, it is difficult to design a universal xml to json parser. As a result, it is challenging to manually create an exhaustive list of structured texts, which are appropriate for the mining process.
2. Unlike the open-access articles, the mining of articles behind the paywall might represent copyright infringement risks.

## CONCLUSIONS

The muddy_db represents the first biologically oriented mud volcano database. It was designed to provide a comprehensive data corpus that can facilitate mud volcano research and shed light on the topic as a whole. The muddy_db contains data ranging from taxonomy to geology and experimental methods. Simultaneously, the muddy_mine NLP pipeline can serve as an example of accessible implementation of NLP techniques in environmental sciences.

## REFERENCES

Bose, A., Rogers, D., Adams, M., Joye, S., and Girguis, P. (2013). Geomicrobiological linkages between short-chain alkane consumption and sulfate reduction rates in seep sediments. *Frontiers in Microbiology*, 4.

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., and Borges, B. (2021). *shiny: Web Application Framework for R*. R package version 1.6.0.

Cheng, T., Chang, Y., Tang, S., Tseng, C., Chiang, P., Chang, K., Sun, C., Chen, Y., Kuo, H., Wang, C., Chu, P., Song, S., Wang, P., and Lin, L. (2012). Metabolic stratification driven by surface and subsurface interactions in a terrestrial mud volcano. *ISME J*, 6(12):2280–2290.

Cui, M., Ma, A., Qi, H., Zhuang, X., and Zhuang, G. (2014). Anaerobic oxidation of methane: an "active" microbial process. *MicrobiologyOpen*, 4(1):1–11.

Etiope, G., Feyzullayev, A., and Baciu, C. (2009). Terrestrial methane seeps and mud volcanoes: A global perspective of gas origin. *Mar Pet Geol*, 26(3):333–344.

Fay, C., Guyader, V., Rochette, S., and Girard, C. (2021). *golem: A Framework for Robust Shiny Applications*. R package version 0.3.1.

Filip, S. and Igras, K. (2021). *semantic.dashboard: Dashboard with Fomantic UI Support for Shiny*. R package version 0.2.0.

Gieg, L., Jack, T., and Foght, J. (2011). Biological souring and mitigation in oil reservoirs. *Applied Microbiology and Biotechnology*, 92(2):263–282.

Honnibal, M. and Johnson, M. (2015). Proceedings of the 2015 conference on empirical methods in natural language processing. Association for Computational Linguistics.

Kleindienst, S., Herbst, F., Stagars, M., von Netzer, F., von Bergen, M., Seifert, J., Peplies, J., Amann, R., Musat, F., Lueders, T., and Knittel, K. (2014). Diverse sulfate-reducing bacteria of the desulfosarcina/desulfococcus clade are the key alkane degraders at marine seeps. *ISME J*, 8(10):2029–2044.

Lo, K., Wang, L., Neumann, M., Kinney, R., and Weld, D. (2020). Proceedings of the 58th annual meeting of the association for computational linguistics. Association for Computational Linguistics.

Luo, J., Wu, M., Gopukumar, D., and Zhao, Y. (2016). Big data application in biomedical research and health care: A literature review. *Biomed Inform Insights*, 8:BII.S31559.

Mazzini, A. and Etiope, G. (2017). Mud volcanism: An updated review. *Earth Sci Rev*, 168:81–112.

Milkov, A. (2000). Worldwide distribution of submarine mud volcanoes and associated gas hydrates. *Mar Geol*, 167(1-2):29–42.

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). Proceedings of the 18th bionlp workshop and shared task. Association for Computational Linguistics.

Pons, M., Quitte, G., Fujii, T., Rosing, M., Reynard, B., Moynier, F., Douchet, C., and Albarede, F. (2011). Early archean serpentine mud volcanoes at isua, greenland, as a niche for early life. *Proc. Natl. Acad. Sci. U.S.A.*, 108(43):17639–17643.

Sano, Y., Kinoshita, N., Kagoshima, T., Takahata, N., Sakata, S., Toki, T., Kawagucci, S., Waseda, A., Lan, T., Wen, H., Chen, A., Lee, H., Yang, T., Zheng, G., Tomonaga, Y., Roulleau, E., and Pinti, D. (2017). Origin of methane-rich natural gas at the west pacific convergent plate boundary. *Scientific Reports*, 7(1).

Schoch, C., Ciufo, S., Domrachev, M., Hotton, C., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J., Sun, L., Turner, S., and Karsch-Mizrachi, I. (2020). Ncbi taxonomy: a comprehensive update on curation, resources and tools. *Database*, 2020.

Vandenbroucke, M. and Largeau, C. (2007). Kerogen origin, evolution and structure. *Org Geochem*, 38(5):719–833.

Wang, J., Deng, H., Liu, B., Hu, A., Liang, J., Fan, L., Zheng, X., Wang, T., and Lei, J. (2020). Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: Bibliometric study on pubmed. *J Med Internet Res*, 22(1):e16816.

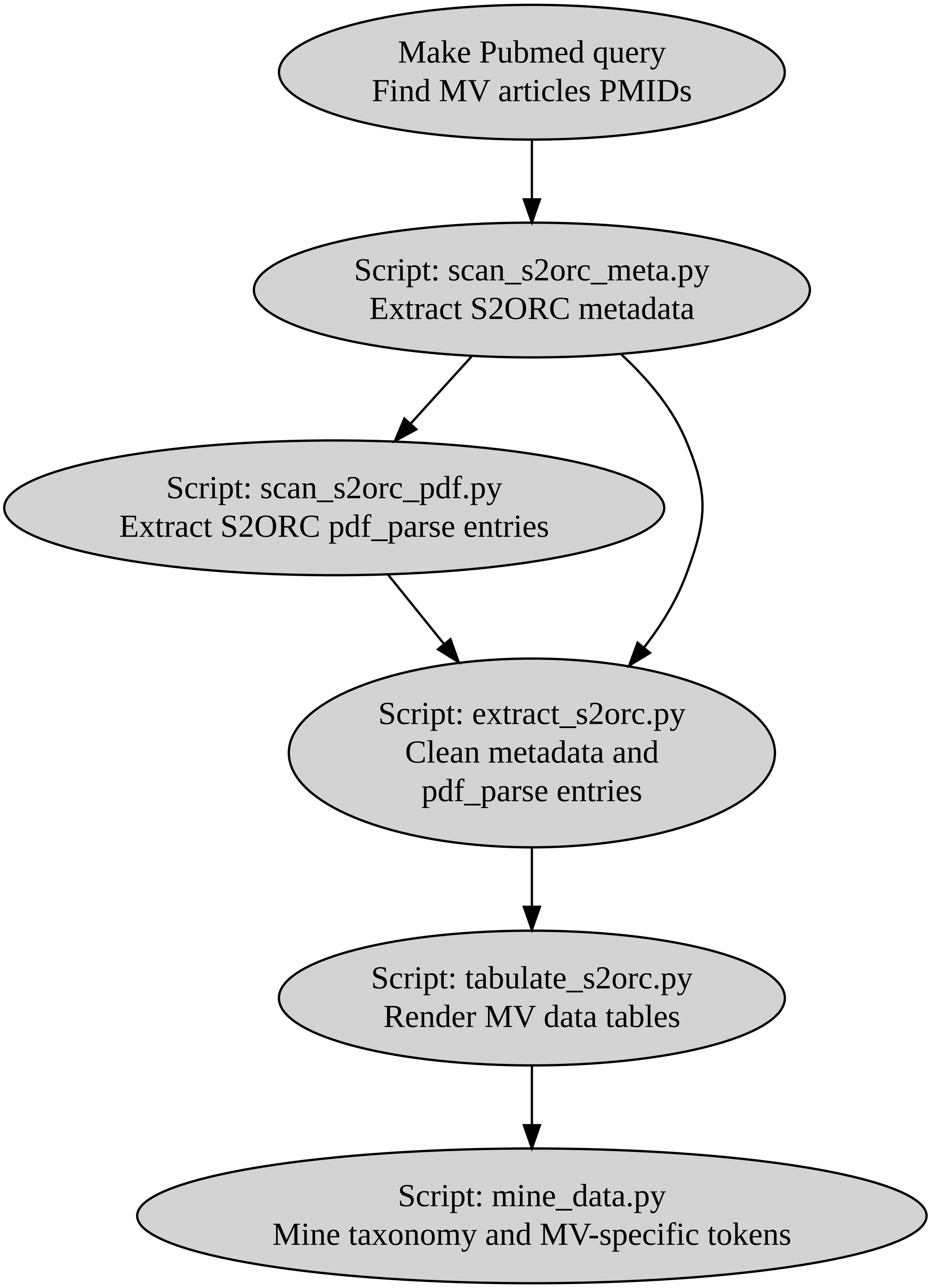
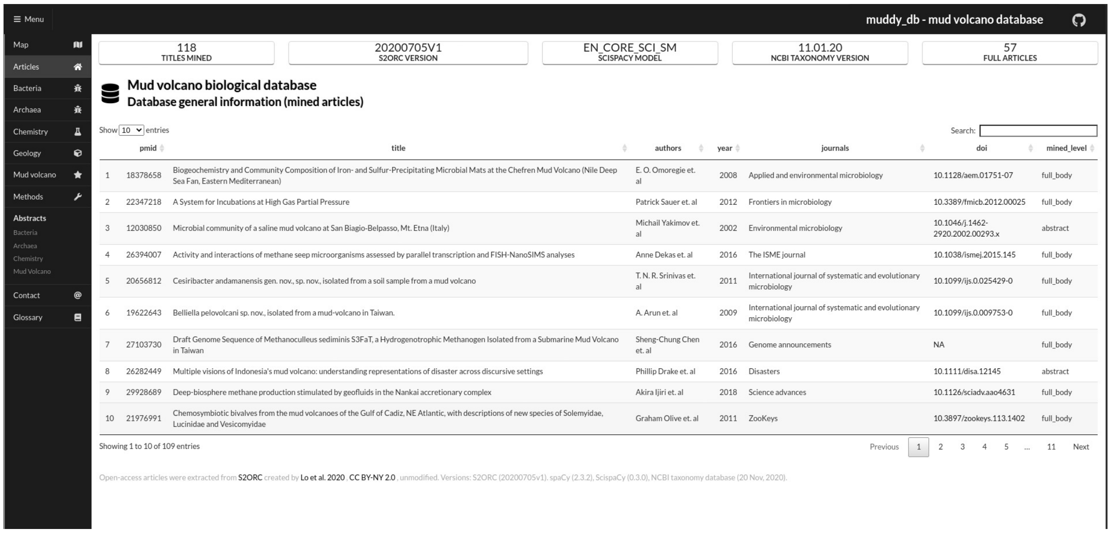


**Figure 1.** muddy_mine - pipeline used to build the muddy_db database. MV - mud volcano, PMIDs - Pubmed



muddy_db - mud volcano database

| 118 | 20200705V1 | EN_CORE_SCI_SM | 11.01.20 | 57 |
|---|---|---|---|---|
| TITLES MINED | S2ORC VERSION | SCISPACY MODEL | NCBI TAXONOMY VERSION | FULL ARTICLES |

**Mud volcano biological database**
**Database general information (mined articles)**

Show 10 entries

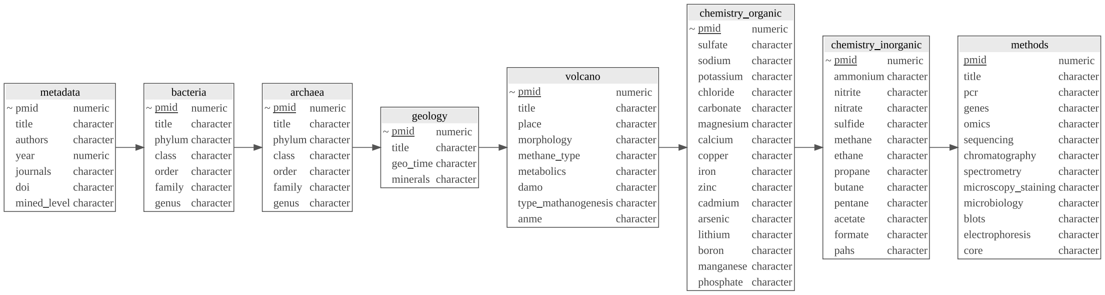
Search:

| | pmid | title | authors | year | journals | doi | mined_level |
|---|---|---|---|---|---|---|---|
| 1 | 18378658 | Biogeochemistry and Community Composition of Iron- and Sulfur-Precipitating Microbial Mats at the Chefren Mud Volcano (Nile Deep Sea Fan, Eastern Mediterranean) | E. O. Omoregie et. al | 2008 | Applied and environmental microbiology | 10.1128/aem.01751-07 | full_body |
| 2 | 22347218 | A System for Incubations at High Gas Partial Pressure | Patrick Sauer et. al | 2012 | Frontiers in microbiology | 10.3389/fmicb.2012.00025 | full_body |
| 3 | 12030850 | Microbial community of a saline mud volcano at San Biagio-Belpasso, Mt. Etna (Italy) | Michail Yakimov et. al | 2002 | Environmental microbiology | 10.1046/j.1462-2920.2002.00293.x | abstract |
| 4 | 26394007 | Activity and interactions of methane seep microorganisms assessed by parallel transcription and FISH-NanoSIMS analyses | Anne Dekas et. al | 2016 | The ISME journal | 10.1038/ismej.2015.145 | full_body |
| 5 | 20656812 | Cesiribacter andamanensis gen. nov., sp. nov., isolated from a soil sample from a mud volcano | T. N. R. Srinivas et. al | 2011 | International journal of systematic and evolutionary microbiology | 10.1099/ijs.0.025429-0 | full_body |
| 6 | 19622643 | Belliella pelovolcani sp. nov., isolated from a mud-volcano in Taiwan. | A. Arun et. al | 2009 | International journal of systematic and evolutionary microbiology | 10.1099/ijs.0.009753-0 | full_body |
| 7 | 27103730 | Draft Genome Sequence of Methanoculleus sediminis S3FaT, a Hydrogenotrophic Methanogen Isolated from a Submarine Mud Volcano in Taiwan | Sheng-Chung Chen et. al | 2016 | Genome announcements | NA | full_body |
| 8 | 26282449 | Multiple visions of Indonesia's mud volcano: understanding representations of disaster across discursive settings | Phillip Drake et. al | 2016 | Disasters | 10.1111/disa.12145 | abstract |
| 9 | 29928689 | Deep-biosphere methane production stimulated by geofluids in the Nankai accretionary complex | Akira Ijiri et. al | 2018 | Science advances | 10.1126/sciadv.aao4631 | full_body |
| 10 | 21976991 | Chemosymbiotic bivalves from the mud volcanoes of the Gulf of Cadiz, NE Atlantic, with descriptions of new species of Solemyidae, Lucinidae and Vesicomyidae | Graham Olive et. al | 2011 | ZooKeys | 10.3897/zookeys.113.1402 | full_body |

Showing 1 to 10 of 109 entries

Open-access articles were extracted from S2ORC created by Lo et al. 2020 , CC BY-NY 2.0 , unmodified. Versions: S2ORC (20200705v1). spaCy (2.3.2), ScispaCy (0.3.0), NCBI taxonomy database (20 Nov, 2020).

**Figure 2.** muddy_db general appearance



| metadata | |
|---|---|
| ~ pmid | numeric |
| title | character |
| authors | character |
| year | numeric |
| journals | character |
| doi | character |
| mined_level | character |

| bacteria | |
|---|---|
| ~ pmid | numeric |
| title | character |
| phylum | character |
| class | character |
| order | character |
| family | character |
| genus | character |

| archaea | |
|---|---|
| ~ pmid | numeric |
| title | character |
| phylum | character |
| class | character |
| order | character |
| family | character |
| genus | character |

| geology | |
|---|---|
| ~ pmid | numeric |
| title | character |
| geo_time | character |
| minerals | character |

| volcano | |
|---|---|
| ~ pmid | numeric |
| title | character |
| place | character |
| morphology | character |
| methane_type | character |
| metabolics | character |
| damo | character |
| type_mathanogenesis | character |
| anme | character |

| chemistry_organic | |
|---|---|
| ~ pmid | numeric |
| sulfate | character |
| sodium | character |
| potassium | character |
| chloride | character |
| carbonate | character |
| magnesium | character |
| calcium | character |
| copper | character |
| iron | character |
| zinc | character |
| cadmium | character |
| arsenic | character |
| lithium | character |
| boron | character |
| manganese | character |
| phosphate | character |

| chemistry_inorganic | |
|---|---|
| ~ pmid | numeric |
| ammonium | character |
| nitrite | character |
| nitrate | character |
| sulfide | character |
| methane | character |
| ethane | character |
| propane | character |
| butane | character |
| pentane | character |
| acetate | character |
| formate | character |
| pahs | character |

| methods | |
|---|---|
| pmid | numeric |
| title | character |
| pcr | character |
| genes | character |
| omics | character |
| sequencing | character |
| chromatography | character |
| spectrometry | character |
| microscopy_staining | character |
| microbiology | character |
| blots | character |
| electrophoresis | character |
| core | character |

**Figure 3.** muddy_db schema