






# Extracting and Modeling Tabular Data from Marine Geology Publications into a Heterogeneous Information Network

Muhammad Asif Suryani<sup>1</sup><sup>a</sup>, Ewa Burwicz-Galerie<sup>2</sup><sup>b</sup>, Brigitte Mathiak<sup>1</sup><sup>c</sup>,  
Klaus Wallmann<sup>3</sup><sup>d</sup> and Matthias Renz<sup>4</sup><sup>e</sup>

<sup>1</sup>*GESIS - Leibniz-Institute for the Social Sciences, 50667 Cologne, Germany*

<sup>2</sup>*MARUM - Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany*

<sup>3</sup>*GEOMAR Helmholtz Centre for Ocean Research Kiel, 24148 Kiel, Germany*

<sup>4</sup>*Institute of Informatik, Christian-Albrechts-Universität zu Kiel, 24118 Kiel, Germany*

{*asif.suryani, brigitte.mathiak*}@*gesis.org*, *eburwicz-galerie@marum.de*, *kwallmann@geomar.de*,

**Keywords:** Information Extraction, Tabular Data, Research Data Management, Marine Science Publication, Heterogeneous Information Network, Data Modeling.


**Abstract:** Scientific publications serve as a source of disseminating information across research communities, often containing diverse data elements such as plain-text, tables, and figures. Tables in particular offer a structured presentation of essential research data, enabling efficient information access. Automatic extraction of tabular data alongside contextual information from scientific publications can significantly enhance research workflows and integrate more research data into scholarly research cycle, particularly supporting Research Data Management (RDM). In marine geology, the researchers conduct expeditions at oceanographic locations and accumulate substantial amounts of valuable data such as Sedimentation Rate (SR), Mass Accumulation Rate (MAR) alongside relevant contextual information, often enriched with spatio-temporal context in tables of publications. These expeditions are costly and time intensive, emphasizing on the value of making such data more accessible and reusable. This paper introduces an end to end approach to extract and model heterogeneous tabular data from marine geology publications. Our approach extracts metadata and tabular content from publications, modeling them into a Heterogeneous Information Network (HIN). The network uncovers hidden relationships and patterns across multiple documents, offering new insights and facilitating enhanced data referencing. Experimental results and exploration on marine geology datasets demonstrate the effectiveness of our approach, showcasing its potential to support research data management and data driven scientific exploration.


## 1 INTRODUCTION


Scientific publications are valuable sources of information that presents data in plain text, tables, and figures. These publications are generally available in Portable Document Format (PDF). These PDFs provide access to the relevant information in a streamlined reading experience. However, due to the rapid increase in the number of publications, the manual extraction of relevant information is becoming dif-


ficult and time-consuming. Moreover, the complex internal structure of PDFs also makes information extraction from plain text, tables, and figures even more challenging. PDFs are primarily designed for platform-independent viewing and printing (Petersen et al., 2021).


To automatically extract the relevant information from these data components of publications it is necessary to have a broader information extraction coverage. As for the context elaboration, researcher utilize plain-text and for graphical representation images are presented. However, tables are generally showcased as a set of organized information which researchers want to emphasize in their research studies. For example numerical information is generally organized in tables and relatively easily accessible to users. But

<sup>a</sup>  <https://orcid.org/0000-0003-1669-5524>

<sup>b</sup>  <https://orcid.org/0000-0003-4551-5609>

<sup>c</sup>  <https://orcid.org/0000-0003-1793-9615>

<sup>d</sup>  <https://orcid.org/0000-0002-1795-376X>

<sup>e</sup>  <https://orcid.org/0000-0002-2024-7700>

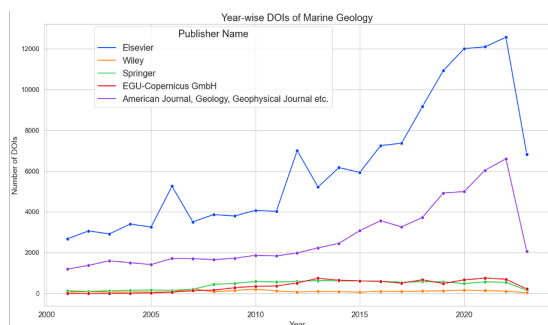


Figure 1: Publications by Years (Hendricks et al., 2020).

the access also depends on the granularity of the information available in these tables (Göpfert et al., 2022).

Moreover, a broader extraction approach is getting indispensable which could support the extraction of complex quantitative information from these tables of publications to overcome manual procedures of information extraction. Besides, there are various factors which need to be considered in this regard, as tables are structured quite diversely regarding the number of (nested) rows and columns, table orientation and depend on the template of publishers (Martinez-Rodriguez et al., 2020),(Ceritli and Williams, 2021).

Marine Geologists perform extensive experiments to record diverse scientific measurements at various oceanographic locations. These experiments are quite expensive in terms of efforts and cost. As researchers have to spend weeks on expedition ships under challenging weather conditions to collect the samples to study the marine environment. There have been various parameters which could be of interest of marine geologists such as the Sedimentation Rate (SR) and Mass Accumulation Rate (MAR) at the seafloor is an excellent example (Chuang et al., 2019). Marine geologists have recorded these measurements for over one hundred years and typically reported them in tables of respective scientific publications. So, proper representation of such quantitative information would be tremendously helpful for both readers and also for the automatic extraction process to carry out relevant extraction conveniently. In order to further emphasis on the importance and scale of this study the information on the number of publications is presented in Figure 1 (Petersen et al., 2021; Göpfert et al., 2022),(Martinez-Rodriguez et al., 2020),(Ceritli and Williams, 2021), (Hendricks et al., 2020).

To further emphasis on the tabular representations of targeted measurements i.e. Sedimentation Rate and Mass Accumulation Rate, it is essential to explain their relevant contextual information. In Figure 2 a sample table comprise of various columns inciting contextual information. The most important features are: *Core* which specifies the name

Region	Core	Lat °N	Long °E	Water depth (m)	Average Sedimentation rate (cm/Kyr)	Proxy	Reference
East Subarctic	AT26-19-05PC	44.973	229.122	2711	0.93	<sup>231</sup> Pa/ <sup>210</sup> Pb	this study
East Subarctic	AT26-19-09PC/06MC	44.887	229.363	2678	1.95	<sup>231</sup> Pa/ <sup>210</sup> Pb	this study
East Subarctic	AT26-19-12PCTC/10MC	44.898	229.496	2689	0.63	<sup>231</sup> Pa/ <sup>210</sup> Pb	this study
East Subarctic	AT26-19-35PC	44.991	229.543	2731	1.85	<sup>231</sup> Pa/ <sup>210</sup> Pb	this study
East Subarctic	AT26-19-38PC	44.971	229.393	2655	1.11	<sup>231</sup> Pa/ <sup>210</sup> Pb	this study
East Subarctic	AT26-19-398B	45.045	229.167	2794	1.18	<sup>231</sup> Pa/ <sup>210</sup> Pb	this study
East Subarctic	W8709A-8	42.270	232.320	3111	9.49	Organic C	Kienast (2003)
East Subarctic	W8709A-13	42.120	234.250	2712	14.11	Organic C	Kienast (2003)
West Subarctic	02P982	50.330	169.500	3244	6.03	Biogenic Ba	Jaccard et al. (2009)
East Subarctic	02P987	54.220	211.730	3647	7.08	Biogenic Ba	McDonald et al. (1999)
West Subarctic	8NDP-PC-13	49.720	168.300	2993	4.33	Biogenic Ba	Brunelle et al. (2010)
Okhotsk Sea	GGC27	49.601	150.180	995	2.52	Biogenic Ba	Brunelle et al. (2010)
Bering Sea	JPC17	53.933	178.699	2209	12.90	Biogenic Ba	Brunelle et al. (2007)
Bering Sea	U1342	54.828	176.917	818	3.65	Organic C	Knudson and Ravelo (2015a)
Okhotsk Sea	PC936	51.015	148.313	1305	5.35	Organic C	Gorbatenko et al. (2004)
West Subarctic	MR98-05-3PC	50.000	164.983	5507	3.59	Biogenic Ba	Shigemitsu et al. (2007)
Okhotsk Sea	X98-01PC	51.015	152.008	1100	8.11	Biogenic Ba	Sato et al. (2002)
Bering Sea	SO201-2-85	57.505	170.413	975	10.67	Organic C	Riebethdorf, Nurnberg, et al. (2013)
Bering Sea	SO201-2-77	56.330	170.699	2133	9.49	Organic C	Riebethdorf, Nurnberg, et al. (2013)

Figure 2: Sample Table taken from PDF (Costa et al., 2018).

of location where the experimentation has been performed. *Lat/Long* columns provide the location information about these measurements. Similarly *Water Depth* provide the depth information. *Average Sedimentation rate* column represents the values being measured at these specific locations. Lastly *Proxy* showcases which method is being adopted to record these measurements. Moreover, all the features are elaborated in order of their contextual importance. It is essential to note that the header representations for these features vary widely across tables, and the targeted measurements do not have proper representations in the International System of Units (SI), making the extraction process even more challenging (Costa et al., 2018).

Extracting relevant information from these table instances can address a wide range of research questions including:

1. Where were these measurements taken, and what are the identifiers for these locations?
2. What are the reported values and units of SR and MAR?
3. Which authors reported these values?
4. What are the connections between studies based on authors and locations?
5. What are the methods and depths associated with these measurements?

However, an access to such information alongside their respective contextual information would provide marine geologists with broader insights and benefit them by having convenient access to their desired information and could accelerate the knowledge discovery process (Suryani et al., 2022). It is also worth highlighting that these targeted measurements and their corresponding spatial information have never been compiled in any repository yet but were reported primarily in publications. This paper discusses the

framework that extracts tables from scientific publications parse these tables to extract quantitative information from marine geology publications to extract relevant information. Finally, the extracted information alongside respective metadata is being exploited to populate a Heterogeneous Information Network comprising of metadata and measurements from the respective publications.

## 2 RELATED WORK

Scientific publications play a crucial role in research by presenting detailed information about experiments and results. Thus access to such research data could be beneficial for the communities to define future research directions. However, research data in this problem setting may provide specialized numerical and spatial information, which facilitates the information acquisition process about experiments or expeditions. The primary function of Research Data Management (RDM) is to provide essential information about conducted experiments, showcasing previously relevant and interconnected data. This contributes to advancing FAIR data principles and improving the delivery of research data across communities (Ducatteuw, 2021).

Marine Geologists mainly report scientific measurements such as Mass Accumulation Rate (MAR) and Sedimentation Rate (SR) in publications, represented in various data elements such as text and tables. These documents do not support a fully automatic information extraction from these data elements of publications. Contrarily, numerous studies previously focused on information extraction from these data elements of publications but only to certain extent (Suryani et al., 2022).

Metadata extraction has evolved over time and various approaches are available, the most prominent being Grobid, which is used in numerous studies (Lopez, 2009). Other metadata-related studies have focused on collaboration networks, citation analysis and publication trends (Moulin and Amaral, 2020; Wahle et al., 2022).

Tabula is an open-source library that allows users to extract tables from PDFs into CSVs (tab, 2022). Camelot is another open-source Python library which extracts tables from PDFs in CSVs. In addition, camelot offers modularity and allows to make it adaptable in any extraction pipeline (cam, 2022).

Recently, measurement extraction from tabular data was carried out using Quantulum3 module, a community-maintained Python library, which focuses on extraction of volumetric units, i.e. liter and pints

from the tabular data. However, the extraction was only performed on spreadsheets instead of PDFs (Ceritli and Williams, 2021). ExtracTable is an approach recently proposed that extracts tabular data from text files and CSVs by detecting the row patterns and separating the columns accordingly (Hübscher et al., 2023). However, this approach not able to address table extraction from the PDFs.

Recently authors proposed Data Acquisition Framework (DAF) which extracts diverse data elements from scientific publications. The DAF framework has ability to extract plain-text, tables, images and metadata from scientific publications with minimal template dependency. The DAF exploits hybrid approaches to perform the extraction from publications and performed better on chemical domain and marine science publications (Suryani et al., 2023), (Zhu and Cole, 2022).

Moreover, the network modeling of diverse information into a network could be challenging as well as interesting and previously addressed numerous applications. In a recent study a novel approach based on Heterogeneous Information Network (HIN) targets potential relationships such as citation links, author collaborations, and research areas. By populating such network using a random walk strategy to simulate natural sentences, the approach effectively discovers relevance between papers (Du et al., 2020).

With the rapid growth of digital publishing, efficiently visualizing scholarly data has become increasingly demanding. This data includes millions of raw data points such as authors, papers, citations, and scholarly networks. Various visualization techniques can be applied to better represent data structures and uncover hidden patterns. The study introduces the basic concepts and collection methods for scholarly data and provides a comprehensive overview of related visualization tools and techniques (Liu et al., 2022).

## 3 TABLE TO NETWORK

The section describes the respective framework and discusses its individual components. The framework consists of Tabular Data Extraction, Information Parser and Network Modeling modules as depicted in Figure 3. Generally, PDF is the primary format in which scientific publications are available, so for extraction of various data elements the focus will be on the raw PDF documents.

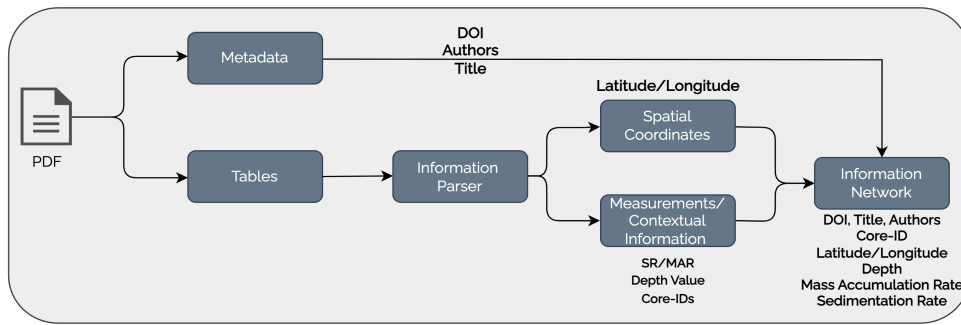


Figure 3: Block Diagram.

### 3.1 Data Extraction

The Data Extraction module takes on scientific publications to perform two crucial tasks, i.e. metadata extraction and table extraction. To extract these data components, we exploited the implementation of our previously proposed framework i.e. Data Acquisition Framework (DAF) (Suryani et al., 2023). The framework perform hybrid approach to extract the metadata features such as DOI, Authors and Title. The metadata features are generally available on the first page of publication and follows certain patterns. The DAF also utilizes various routines to acquire metadata features and later provide a consolidated set of metadata features. For tabular data, DAF breaks the task into two distinct steps i.e. Detection and Extraction. For the detection of tables it transform the pages of the PDF document to coordinates and by heuristic of caption of tables to find the bounding box of the prospective tabular regions and later passed these coordinates to base (cam, 2022) module for the extraction of tabular data to comma separated files. It is also important to mention that DAF has better extraction coverage of metadata features targeting domain specific research publications as compared to (Zhu and Cole, 2022) in chemical domain publications. The extracted sample of tabular data is shown in Figure 4.

### 3.2 Information Parser

The Information Parser module exploits the extracted tabular data from data extraction module. Information parser module plays a crucial role in this problem setting by initially performing table processing, which involves information detection, correction and segregation tasks. The information parser module is specifically designed to handle tables within the context of marine geology publications, considering these essential attributes:

1. Distinct identifiers associated with each oceanographic location such as Core-ID.

Region	Core	Lat °N	Long °E	Water depth (m)	Average Sedimentation rate (m/kyr)	Proxy	Reference
East Subarctic	A726-19-05PC	44.973	229.122	2711	0.93	231Pa/230Th	this study
East Subarctic	A726-19-09PC/08MC	44.887	229.363	2678	1.95	231Pa/230Th	this study
East Subarctic	A726-19-12PC-TIC/10MC	44.898	229.496	2689	0.83	231Pa/230Th	this study
East Subarctic	A726-19-35PC	44.991	229.543	2731	1.85	231Pa/230Th	this study
East Subarctic	A726-19-38PC	44.971	229.393	2655	1.11	231Pa/230Th	this study
East Subarctic	A726-19-38BB	45.045	229.167	2794	1.18	231Pa/230Th	this study
East Subarctic	W8709A-8	42.270	232.320	3111	0.49	Organic C	Kienast (2003)
East Subarctic	W8709A-13	42.120	234.250	2712	14.11	Organic C	Kienast (2003)
West Subarctic	ODP882	50.330	167.500	3244	6.03	Biogenic Ba	Jaccard et al. (2009)
East Subarctic	ODP887	54.220	211.730	3647	7.08	Biogenic Ba	McDonald et al. (1999)
West Subarctic	RNDP-PC-13	49.720	168.300	2383	4.33	Biogenic Ba	Brunelle et al. (2010)
Okhotsk Sea	GSC27	49.601	150.180	995	2.52	Biogenic Ba	Brunelle et al. (2010)
Bering Sea	JPC17	53.933	178.699	2209	12.90	Biogenic Ba	Brunelle et al. (2007)
Bering Sea	U1342	54.828	176.917	818	3.05	Organic C	Knudsen and Revuelto (2015a)
Okhotsk Sea	PC936	51.015	148.313	1305	5.35	Organic C	Gorbanenko et al. (2004)
West Subarctic	MR98-05-3PC	50.000	164.983	5507	3.59	Biogenic Ba	Shigemitsu et al. (2007)
Okhotsk Sea	X98-01PC	51.015	152.008	1100	8.11	Biogenic Ba	Sato et al. (2002)
Bering Sea	SO201-2-85	57.505	170.413	975	10.67	Organic C	Riehdorf, Nurnberg, et al. (2013)
Bering Sea	SO201-2-77	56.300	170.699	2133	8.49	Organic C	Riehdorf, Nurnberg, et al. (2013)

Figure 4: Example of Extracted Table by DAF (Costa et al., 2018).

2. Latitude and Longitude: Expressed in either degrees or decimals.
3. Mass Accumulation Rate (MAR) and Sedimentation Rate (SR) expressed in *Mass/Area/Time* and *Length/Time* respectively.
4. Water Depth values offering vertical position to respective identifiers generally expressed in meters.

To address these information segregation from tables, we studied over 100 relevant papers to have an overview of possible information representation scenarios for each of the targeted features. We carefully compiled an extensive dictionary encompassing a multitude of potential variants for each relevant feature. Particularly, we identified twelve distinct variants for Core-IDs, fifteen variations for location data, sixteen alternatives for MAR and SR headers, and seven diverse representations for water depth measurements.

The Figure 4 shows the extracted table from the relevant publications, which indicates the efficacy of our DAF for the extraction of tabular data from pub-

Region	Water depth (m)	rate (cm/kyr)	Proxy	Coordinates
East Subarctic	2711	0.93	231Pa/230Th	44.973N,229.122E
East Subarctic	2678	1.95	231Pa/230Th	44.887N,229.363E
East Subarctic	2689	0.63	231Pa/230Th	44.898N,229.496E
East Subarctic	2731	1.85	231Pa/230Th	44.991N,229.543E
East Subarctic	2655	1.11	231Pa/230Th	44.971N,229.393E
East Subarctic	2794	1.18	231Pa/230Th	45.045N,229.167E
East Subarctic	3111	9.49	Organic C	42.270N,232.320E
East Subarctic	2712	14.11	Organic C	42.120N,234.250E
West Subarctic	3244	6.03	Biogenic Ba	50.330N,167.500E
East Subarctic	3647	7.08	Biogenic Ba	54.220N,211.730E
West Subarctic	2393	4.33	Biogenic Ba	49.720N,168.300E
Okhotsk Sea	995	2.52	Biogenic Ba	49.601N,150.180E
Bering Sea	2209	12.90	Biogenic Ba	53.933N,178.699E
Bering Sea	818	3.65	Organic C	54.828N,176.917E
Okhotsk Sea	1305	5.35	Organic C	51.015N,148.313E
West Subarctic	5507	3.59	Biogenic Ba	50.000N,164.983E
Okhotsk Sea	1100	8.11	Biogenic Ba	51.015N,152.008E
Bering Sea	975	10.67	Organic C	57.505N,170.413E
Bering Sea	2133	9.49	Organic C	56.330N,170.699E

Figure 5: Parsed Table Example (Costa et al., 2018).

lications. Later, each extracted table is parsed to extract relevant features while maintaining internal relationships intact. Core-IDs often lack a fixed pattern, encompassing alphanumeric strings, numbers, or arbitrary names. So for Core-IDs, the parser references our dictionary and tracks the relevant column. A similar approach is employed for identifying depth columns, though their presence is not usual. However, parsing latitude and longitude information from extracted tables is the challenging task. As oceanographic location coordinates are expressed in degrees and minutes encounter encoding issues, such as “ ° ” becoming “ 8 ” or “ 0 ” and “ ’ ” turning into “ 1 ”, such complexities are not much in numbers but are complex and not handled in this study. The sample of parsed table is shown in Figure 5. Additionally, measurement columns like MAR and SR are cross-referenced with dedicated headers and units dictionaries respectively to have relevant measurements and their corresponding location coordinates and core-ids.

The Figure 4 showcases the parsed tabular information performed by the information parser which indicates the ability to reproduce the tabular content from the marine geology publications (Costa et al., 2018). Hence indicating the importance and pave the way to have tabular research data from relevant scientific publications available across research.

### 3.3 Network Modeling

This section explains the proposed data model which exploits metadata features and parsed tabular information from the publications. Heterogeneous Information Network (HIN) is used to represent and analyze complex linked data comprising various types of entities and relationships. HIN incorporates diverse node and edge types, capturing the multifaceted na-

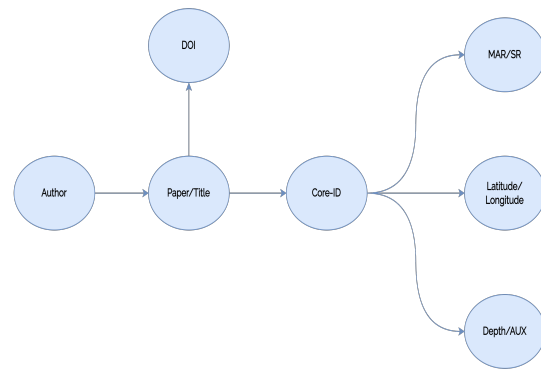


Figure 6: Proposed Data Model.

ture of diverse data. HIN can be defined as a graph  $G = (V, E)$  where  $V$  denotes the set of nodes (or vertices) and  $E$  represents the set of edges. This representation and understanding of interconnected data, makes HINs particularly useful in various potential studies (Liu et al., 2022). In this problem setting, HINs are populated by leveraging the proposed data models, which integrate the relational data and metadata by performing exploratory analyses to showcase the broader spectrum of information from publications.

The data model leverages the extracted metadata features alongside the parsed tabular data from the respective publication sources. Figure 6 illustrates the data model, which organizes heterogeneous features from marine geology publications. It structures both metadata and tabular data, while capturing document level relationships intact. Furthermore, the data model facilitates the generation of a heterogeneous information network, revealing potential relationships among targeted entities and providing a broader context across documents.

## 4 EXPERIMENTAL RESULTS

This section provides explanation of the experimental setup employed, followed by a brief discussion on the gathered exploratory results.

### 4.1 Data Description

In this exploratory study, publications from Marine Geology were gathered from various publishers. A comprehensive exploration covering over 300 full articles was conducted, serving various purposes, such as collection of diverse expressions of information representation. The selection criteria for these publications was centered around an important require-

Table 1: Collected Results of Metadata Features from Publications.

Metadata Feature	P	R	F1-Score
DOI	1.0	0.90	0.95
Title	0.95	1.0	0.97
Author	1.0	0.90	0.95

ment, as each publication must carry location coordinates and relevant measurements in tables.

Moreover, it is also important to mention that the problem, we are addressing here is not explored before, so to search for the relevant sources require considerable efforts. Furthermore, an evaluation of the overall framework was undertaken on forty full papers from Marine Geology covering different publishers. However, collection of data for experimentation from various publishers will also be important to test the template coverage of Data Acquisition Framework (Suryani et al., 2023).

## 4.2 Results and Discussion

The set of publications are initially processed by Data Acquisition Framework (DAF) and results are collected (Suryani et al., 2023). The first set of results comprise of targeted metadata features such as DOIs, Author information and the title of the publications which will be helpful in answering the questions related to the origination of research activity such as the authors being involved in the research study. Beside, it is also helpful in studying the relationships across metadata features. The results in this regard are compiled and shown in Table 1.

There were a total of 111 table instances across the publications, out of which DAF successfully extracted 106. The criterion for true table extraction was to ensure that all rows and columns remained intact. Among these 106 table instances, 46 were identified as relevant, carrying the desired information for our proposed study. Furthermore, the information parser successfully parsed 40 of these relevant tables, while the remaining 6 encountered issues where location coordinates were misaligned, as described in Section 3.2. The later steps involve modeling the extracted information into a heterogeneous information network. For the sake of simplicity, information modeling is performed at two levels: metadata and complete. By exploiting the importance of metadata features, information modeling is initially performed on the extracted metadata features, showcasing the meta-information to relevant research communities. Figure 7 illustrates a network of metadata features of publications. For better understanding, all node types follow a distinct color scheme, DOI instances are represented

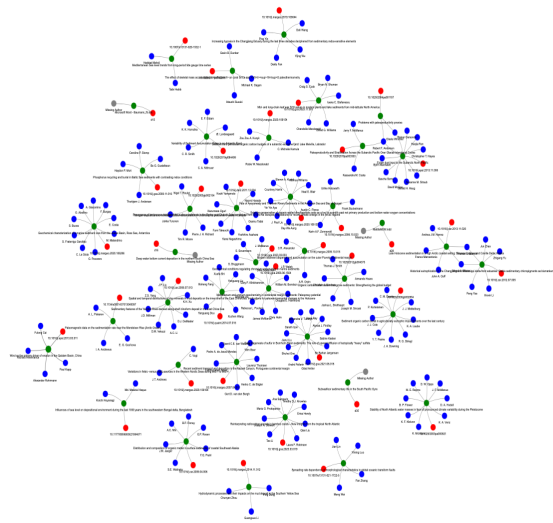


Figure 7: Network based on Metadata of Marine Geology Publications.

in red, author nodes in blue, and titles in green.

Finally, we constructed a comprehensive network which models all the extracted information from the publications i.e. metadata features, relevant measurements and contextual information. For better exploration the features are denoted in different color scheme to enhance the information accessibility. The network in this regard is shown in Figure 8. Here the information against core ID is presented in pink and all relevant tabular information is shown in purple. Moreover, the network in Figure 8 also showcase the author level information as well as the core level connectivity to showcase the diverse relationships. Hence capable of addressing various questions including:

Finally, we constructed a comprehensive network that models all the extracted information from the publications, including metadata features, relevant measurements, and contextual information. To facilitate better exploration, the features are represented using a distinct color scheme to enhance information accessibility as shown in Figure 8.

In this figure, information associated with the core-ID is presented in pink, while all relevant tabular information is displayed in purple. Moreover, the network in Figure 8 also illustrates author-level information and core-level connectivity, highlighting the diverse relationships which address various questions, including:

1. What are the identifiers for the targeted locations?
2. Who measured these values?
3. Where are SR and MAR values measured.
4. What is the water depth value corresponding to these measurements in specific regions?

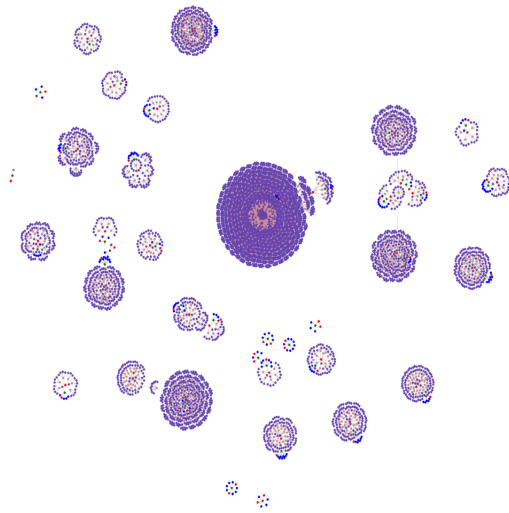


Figure 8: Network Comprised of metadata and parsed tabular data from Marine Geology Publications.

5. Who are the potential collaborators targeting geographical regions?

Hence, to enhance information delivery, a small subset of the complete network is presented in Figure 9, illustrating author level relationships between two documents. Additionally, it highlights potential location based research collaborations and identifies groups or individuals working at specific oceanographic locations.

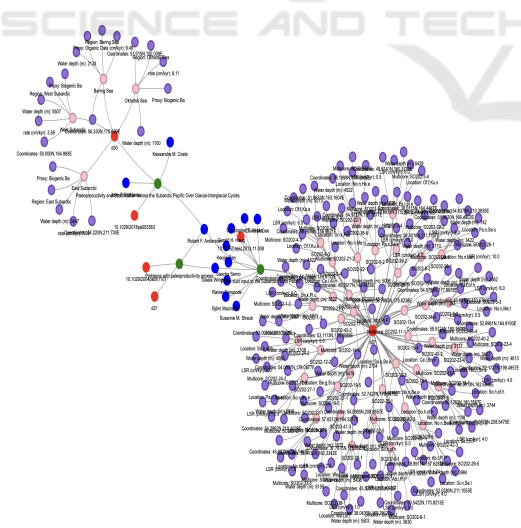


Figure 9: Subset of the Network Showing Potential Relationships across Publications.

This exploratory study aims to highlight the diverse information available across research publications by facilitating research data management and the open science initiative. One of the major aspects is

to integrate more research data from marine geology into the scholarly eco-system at different granularities. Figure 7 aims to provide author level relationships within a set of publications, which are equally essential for an abstract overview.

## 5 CONCLUSION

The proposed framework addresses the challenging task of extracting tables from scientific publications to enhance the information discovery process by retrieving diverse quantitative data along with contextual information from tables in PDF files. This approach contributes to making research data more accessible and usable within the marine geology domain. Unlike prevalent approaches that primarily focus on textual data, our framework emphasizes tabular data, generating valuable insights that were previously limited to the researchers. The potential outcome of the framework is the creation of document-level Heterogeneous Information Network that leverages heterogeneous metadata and tabular content.

In addition, it is essential to discuss the potential challenges in this problem setting. The variance in the representation of tabular data in publications is a key aspect to consider, as it elevates the challenges in the table detection and extraction process. Furthermore, metadata plays a crucial role in facilitating information profiling, however, its extraction is equally challenging, primarily due to the variety of templates. Furthermore, the lack of standardized guidelines for representing such diverse data in marine geology publications is notable and need to be focused in future for a streamlined delivery of information.

Moreover, it is essential to highlight two key aspects: the availability of relevant publications and the extraction of data from these publications. Regarding availability, finding relevant publications while adhering to FAIR principles is challenging, despite the availability of large number of publications. Besides, in terms of data extraction, numerous historical documents exist that are not true PDF files, indicating towards an enhanced data extraction approach capable of extracting relevant information from images.

Beyond marine geology, the proposed framework holds promise as a versatile solution applicable across various scientific domains. Additionally, integrating it with textual information extraction could enable a comprehensive automated approach for efficiently extracting heterogeneous data from all components of scientific publications. Future directions include leveraging the extracted information for applications such as recommender systems, research community

detection, and spatial research collaborations. Furthermore, addressing the diverse representations of tabular data by transforming PDFs into LaTeX expressions represents an exciting future research direction. Lately, the adoption of Large Language Models (LLMs) for various Natural Language Processing (NLP) tasks paves the way for their potential adoption in information extraction, highlighting exciting prospects for future research.

## ACKNOWLEDGEMENTS

This work was partially funded by the Helmholtz Association (grant HIDSS-0005). EB-G received support from the Cluster of Excellence ‘The Ocean Floor – Earth’s Uncharted Interface’ (EXC 2077) funded by Deutsche Forschungsgemeinschaft (DFG) - Project number 390741603 hosted by the Research Faculty MARUM-Center for Marine Environmental Sciences, University of Bremen, Germany. This work has been partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), under SmartER project (Grant number 515537520). Authors also acknowledge the sources being used and the efforts of all collaborators.

## REFERENCES

- (2022). Camelot. Last accessed 4 May 2023.
- (2022). Tabula-py. Last accessed 4 May 2023.
- Ceritli, T. and Williams, C. K. (2021). Identifying the units of measurement in tabular data. *arXiv preprint arXiv:2111.11959*.
- Chuang, P.-C., Yang, T. F., Wallmann, K., Matsumoto, R., Hu, C.-Y., Chen, H.-W., Lin, S., Sun, C.-H., Li, H.-C., Wang, Y., et al. (2019). Carbon isotope exchange during anaerobic oxidation of methane (aom) in sediments of the northeastern south china sea. *Geochimica et Cosmochimica Acta*, 246:138–155.
- Costa, K. M., McManus, J. F., and Anderson, R. F. (2018). Paleoproductivity and stratification across the subarctic pacific over glacial-interglacial cycles. *Paleoceanography and Paleoclimatology*, 33(9):914–933.
- Du, N., Guo, J., Wu, C. Q., Hou, A., Zhao, Z., and Gan, D. (2020). Recommendation of academic papers based on heterogeneous information networks. In *2020 IEEE/ACS 17th International Conference on Computer Systems and Applications (AICCSA)*, pages 1–6. IEEE.
- Ducatteeuw, V. (2021). Developing an urban gazetteer: A semantic web database for humanities data. In *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Geospatial Humanities*, pages 36–39.
- Göpfert, J., Kuckertz, P., Weinand, J., Kotzur, L., and Stolten, D. (2022). Measurement extraction with natural language processing: A review. *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2191–2215.
- Hendricks, G., Tkaczyk, D., Lin, J., and Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427.
- Hübscher, L., Jiang, L., and Naumann, F. (2023). Extractable: Extracting tables from raw data files. *BTW 2023*.
- Liu, J., Shi, C., Yang, C., Lu, Z., and Philip, S. Y. (2022). A survey on heterogeneous information network based recommender systems: Concepts, methods, applications and resources. *AI Open*, 3:40–57.
- Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, pages 473–474. Springer.
- Martinez-Rodriguez, J. L., Hogan, A., and Lopez-Arevalo, I. (2020). Information extraction meets the semantic web: a survey. *Semantic Web*, 11(2):255–335.
- Moulin, T. C. and Amaral, O. B. (2020). Using collaboration networks to identify authorship dependence in meta-analysis results. *Research Synthesis Methods*, 11(5):655–668.
- Petersen, T., Suryani, M. A., Beth, C., Patel, H., Wallmann, K., and Renz, M. (2021). Geo-quantities: A framework for automatic extraction of measurements and spatial context from scientific documents. In *17th International Symposium on Spatial and Temporal Databases*, pages 166–169.
- Suryani, M. A., Hahne, S., Beth, C., Wallmann, K., and Renz, M. (2023). Daf: Data acquisition framework to support information extraction from scientific publications. In *Proceedings of the 15th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR*, pages 468–476. INSTICC, SciTePress.
- Suryani, M. A., Wölker, Y., Sharma, D., Beth, C., Wallmann, K., and Renz, M. (2022). A framework for extracting scientific measurements and geo-spatial information from scientific literature. In *2022 IEEE 18th International Conference on e-Science (e-Science)*, pages 236–245. IEEE.
- Wahle, J. P., Ruas, T., Mohammad, S. M., and Gipp, B. (2022). D3: A massive dataset of scholarly metadata for analyzing the state of computer science research. *arXiv preprint arXiv:2204.13384*.
- Zhu, M. and Cole, J. M. (2022). Pdfdataextractor: A tool for reading scientific text and interpreting metadata from the typeset literature in the portable document format. *Journal of Chemical Information and Modeling*, 62(7):1633–1643.