Generating Research Data Metadata from Their Accompanying README Files

Kotaro Sekido[1], **Yu Watanabe**[1], **Koichiro Ito**[1], **Shigeki Matsubara**[1,2]
[1]Graduate School of Informatics, Nagoya University
[2]Information Technology Center, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464-8601, Japan
{sekido.kotaro.h0,watanabe.yu.x3}@s.mail.nagoya-u.ac.jp
{ito.koichiro.z5,matsubara.shigeki.z8}@f.mail.nagoya-u.ac.jp

Abstract

Software repositories have conventionally been used for software development. Recently, they have also served as research data repositories. Research data published in such repositories are frequently accompanied by README files; however, the data frequently lack structured metadata. To address this issue, this paper investigates the feasibility of generating research data metadata from their accompanying README files. First, we analyze the occurrence patterns of metadata-related information in README files. The results of this analysis demonstrated that README files could serve as valuable resources for metadata generation. We then performed an experiment on extracting metadata-related information from README files using large language models (LLMs) and evaluated their performance. The experimental results demonstrated that LLMs could extract metadata-related information with high performance.

Keywords: open science, research data management, metadata generation, LLM

1. Introduction

Software repositories, e.g., GitHub[1], have been primarily used for software development. In addition to source code, they typically include README files that describe usage and other details. Recently, software repositories have been increasingly used as platforms to publish research data; thus, they also function as repositories for research data (Koesten et al., 2020). In such cases, README files are frequently provided to describe the overview and usage of the corresponding research data.

To facilitate smooth circulation of research data, the data should be described with rich metadata (Wilkinson et al., 2016) comprising structured fields and their values (as shown in Table 1). The registration of metadata in metadata repositories improves the searchability of research data. However, research data published in software repositories frequently lack metadata (Gesese et al., 2025), which makes it difficult for researchers to discover the research data, even if they are accessible to the public.

In this paper, we investigate the feasibility of generating metadata for research data from the accompanying README files. We expect that the searchability of research data published in software repositories can be improved if metadata can be generated from README files and registered in metadata repositories.

[1]https://github.co.jp/

Field	Value
Title	Helsinki Prosody Corpus
PublicationYear	2019
Creator	Aarne Talman, Antti Suni, ...
Subject	Prosody Prediction
Language	English

Table 1: Partial metadata of the Helsinki Prosody Corpus (Talman et al., 2019).

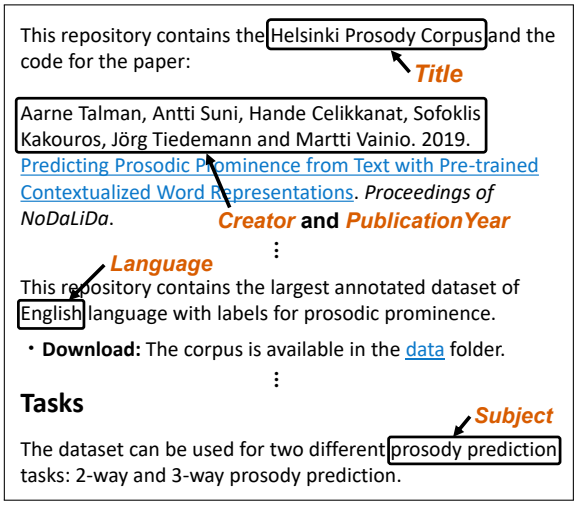

Figure 1: Excerpt from the README file of the Helsinki Prosody Corpus (Talman et al., 2019).

First, we analyzed the occurrence patterns of metadata-related information in README files. In this analysis, we targeted GitHub repositories as

Proceedings of the 3rd Int. Workshop on Natural Scientific Language Processing (NSLP 2026) @ LREC 2026, pages 180–185
12 May 2026.

one of the most widely used software repositories. The results demonstrated that although there were differences among the metadata fields, README files could serve as valuable resources for metadata generation. We then performed an experiment to evaluate the feasibility of extracting metadata-related information from README files. In this experiment, large language models (LLMs) were employed for information extraction, and we evaluated their performance. The results demonstrated that the LLMs could extract metadata-related information from the README files with high performance.

2. Metadata Generation for Research Data

2.1. Research Data Circulation and Metadata

Recently, with global trends toward open science, the importance of publishing and sharing research data has increased (UNESCO, 2021; OSWG, 2023). However, to facilitate the circulation of research data, it is essential to provide rich metadata (Wilkinson et al., 2016). Table 1 shows part of the metadata for the Helsinki Prosody Corpus, which is a dataset used to predict speech prosody, as an example of research data metadata. Research data metadata include information that characterizes the data, e.g., the name of the data (*Title*), year of publication (*PublicationYear*), creators (*Creator*), subject or keywords (*Subject*), and languages (*Language*).

2.2. Metadata Generation using README Files

A README file for research data is a document created to facilitate use of the corresponding data. In other words, README files support human understanding and reuse of the data. Figure 1 shows an excerpt from the README file of the Helsinki Prosody Corpus[2] as a representative example of a README file for research data. A research data README file contains descriptions of the data and may include metadata-related information. For example, the README file shown in Figure 1 includes information corresponding to the metadata presented in Table 1, e.g., the *Title* and *PublicationYear* of the research data. One approach to improving the discoverability of research data in software repositories is to generate metadata from README files and register them in metadata repositories.

[2] https://github.com/Helsinki-NLP/prosody

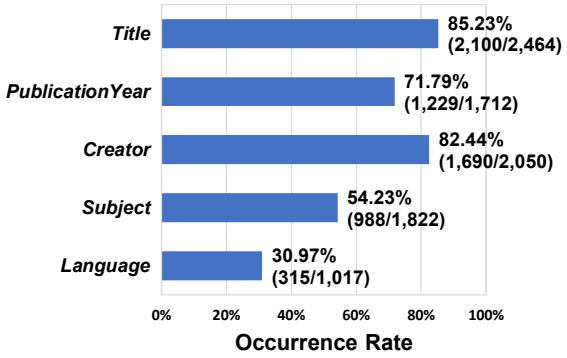

Figure 2: Occurrence rate for each metadata field.

2.3. Related Work

Previous studies have attempted to extract information about research data from documents that mention the data. Most of these studies targeted software or datasets, and they extracted metadata-related information from scholarly papers (Schindler et al., 2021; Stavropoulos et al., 2023; Watanabe et al., 2024; Alyafeai et al., 2025; Watanabe et al., 2025).

Several studies have attempted to acquire information about research data from README files (Kumar et al., 2024; Utrilla Guerrero et al., 2024). These studies focused on software and attempted to extract mentions related to functionality or usage. However, to the best of our knowledge, no previous studies have investigated the use of README files to generate research data metadata.

3. Analysis

In this study, we first evaluated whether README files can serve as valuable resources for metadata generation. Here, we analyzed the occurrence patterns of metadata-related information in README files.

Recently, an increasing number of datasets have been registered in software repositories (Koesten et al., 2020; Roman et al., 2023). Despite this trend, research focusing specifically on such datasets remains limited compared with studies focused on software and other artifacts. Thus, in this analysis, we focused on datasets as the target research data.

3.1. Analysis Data

The analysis required data comprising pairs of metadata and README files for research data. Thus, we constructed the data to be analyzed from Papers With Code Datasets (PWCD) (Papers With Code, 2019–25). PWCD includes dataset metadata and the URLs of their corresponding publication repositories. In this study, we targeted URLs

whose publication platform is GitHub, and we retrieved the README files using the GitHub API. Note that repositories containing multiple README files were excluded from the analysis because it was unclear which README file corresponded to the dataset. In addition, the metadata in PWCD are written in English; thus, we limited our collection to repositories with English README files.

Among the 2,737 collected pairs of metadata and README files, we used 2,464 pairs for analysis, excluding those whose repositories were created after 2023.

Among the metadata fields included in the collected data, we selected those corresponding to fields in the DataCite Metadata Schema (DataCite Metadata Working Group, 2024), which is a global standard metadata schema, as the targets of the analysis. As a result, the five fields shown in Table 1 were selected.

3.2. Analysis Method

The method employed to determine the occurrence of metadata values in the README files is described as follows. First, we applied text preprocessing to both the metadata and README files to reduce the effects of orthographic variations. Then, for each metadata field, we determined whether its value occurred in the README file. In PWCD metadata, the values for *Creator*, *Subject*, and *Language* are represented as lists. For these fields, if at least one element of the list occurred in the README file, we determined that the values for those fields occurred. Note that some metadata fields in PWCD contain missing values. For such cases, these entries were excluded from the calculation of the occurrence rate of the metadata values in README files.

3.3. Analysis Results

Figure 2 shows the calculated occurrence rate for each metadata field. As shown, the fields with the highest occurrence rate were *Title* and *Creator*, both of which exceeded 80%. Furthermore, the field with the lowest occurrence rate, i.e., the *Language* field, reached 30%. The macro-average occurrence rate across the five fields was 64.93%. From these results, although the occurrence rate varied among the fields, we confirmed that README files can serve as valuable resources for metadata generation.

4. Experiment

In this study, we performed an experiment to investigate the feasibility of extracting metadata-related information from README files using LLMs. As in Section 3, we targeted the README files for datasets in this experiment.

```
# System
You are an information extraction model.
Your task is to extract information about dataset from the provided README.
You must output the results strictly according to the following definitions.

Definitions
- name: The name of the dataset.
- introduced_year: The year when the dataset was introduced.
- creators: The creators of the dataset.
- tasks: The tasks that dataset is intended for.
- natural_languages: The natural languages of the dataset.

If a value is not stated in the README, set the value to null.

# User
README:
{readme_text}
```

Figure 3: LLMs prompt.

4.1. Experimental Data

Here, we split the analysis data described in Section 3 into training and development sets, and we allocated data that were not used in the analysis to the test set. The training, development, and test sets included 2,189, 275, and 273 samples, respectively. In this experiment, the development set was used for prompt design, and the test set was used to evaluate the extraction performance.

4.2. Implementation

In this study, we employed Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Ministral-8B-Instruct-2410 (Mistral AI Team, 2024), and Qwen3-8B (Yang et al., 2025) as open LLMs, and GPT-5 (Open AI, 2025) as a closed LLM. Here, we used vLLM[3] (Kwon et al., 2023) for the open LLMs and the OpenAI API[4] for GPT-5. For vLLM, the temperature was set to 0.0. Note that we limited the model responses to JSON format consisting of the five fields shown in Table 1 using Structured Outputs. Figure 3 shows the prompt. The prompt included instructions to extract values for each metadata field and to output "null" if a value could not be identified.

[3] https://github.com/vllm-project/vllm
[4] https://openai.com/ja-JP/api/

	Recall				Precision			
	Llama	Ministral	Qwen	GPT	Llama	Ministral	Qwen	GPT
Title	85.71 (198/231)	84.42 (195/231)	87.45 (202/231)	**91.34 (211/231)**	77.34 (198/256)	79.27 (195/246)	75.94 (202/266)	**83.40 (211/253)**
PublicationYear	80.20 (158/197)	76.14 (150/197)	**93.40 (184/197)**	87.82 (173/197)	88.76 (158/178)	**89.82 (150/167)**	81.78 (184/225)	86.07 (173/201)
Creator	**95.79 (182/190)**	93.68 (178/190)	95.26 (181/190)	87.37 (166/190)	98.38 (182/185)	97.80 (178/182)	89.60 (181/202)	**98.81 (166/168)**
Subject	80.19 (85/106)	75.47 (80/106)	69.81 (74/106)	**91.51 (97/106)**	**55.56 (85/153)**	48.78 (80/164)	40.22 (74/184)	55.43 (97/175)
Language	**74.19 (23/31)**	67.74 (21/31)	67.74 (21/31)	61.29 (19/31)	56.10 (23/41)	46.67 (21/45)	26.58 (21/79)	**67.86 (19/28)**

Table 2: Extraction performance (Recall and Precision).

	Llama	Ministral	Qwen	GPT
Title	81.31	81.76	81.29	**87.19**
PublicationYear	84.27	82.42	**87.20**	86.93
Creator	**97.07**	95.70	92.35	92.74
Subject	65.64	59.26	51.03	**69.04**
Language	63.89	55.26	38.18	**64.41**
Macro average	78.43	74.88	70.01	**80.06**

Table 3: Extraction performance (F1).

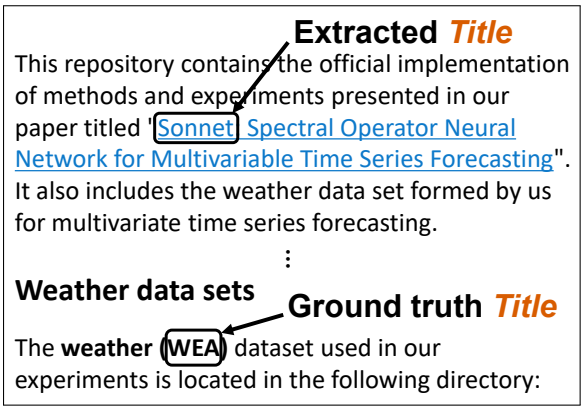

Figure 4: Excerpt from the README file of the WEA (Shu and Lampos, 2025) illustrating an extraction error caused by the presence of multiple research data.

4.3. Evaluation

In this experiment, the extraction performance of the LLMs for each metadata field was evaluated using the recall, precision, and F1-score, which are calculated as follows.

Recall. The proportion of correctly extracted metadata values among those in the README files.

Precision. The proportion of correctly extracted metadata values among non-null values output by the LLMs.

F1-score. The harmonic mean of recall and precision.

We determined whether the extracted values were correct by string matching with the ground truth. Here, partial match was employed for *Title* and *Subject*, and exact match was used for *PublicationYear*, *Creator*, and *Language*. Partial match was defined as a case where either string is a substring of the other. For list-type metadata values, i.e., *Creator*, *Subject*, and *Language*, the extraction was considered correct if at least one element matched. Note that metadata fields with missing values in the test set were excluded from the evaluation.

4.4. Experimental Results

The recall and precision results are shown in Table 2. As can be seen, for recall, although the scores for *Language* were relatively low, the best scores for the other metadata fields exceeded 90%, which indicates that the LLMs could extract metadata values from the README files with high coverage. In terms of precision, although the scores for *Subject* and *Language* were relatively low, the best scores for the other metadata fields exceeded 80%.

Table 3 shows the results for the F1-score. The macro-average for all LLMs falls in the range of approximately 70% to 80%. Among the LLMs, GPT-5 obtained the best performance. These results demonstrated that the LLMs were able to extract metadata-related information from the README files with high performance.

4.5. Error Analysis

Although the LLMs demonstrated high extraction performance, several errors were observed. Specifically, when research data of a different type from the dataset appeared in a README file, the LLMs sometimes incorrectly extracted information related to those nontarget research data. As a representative example, Figure 4 shows an excerpt from the README file of the WEA dataset (Shu and Lampos, 2025)[5] for weather forecasting. Here, al-

[5] https://github.com/ClaudiaShu/Sonnet

though the correct *Title* is “WEA,” “Sonnet,” which is a model trained using “WEA,” was incorrectly extracted as the *Title*.

To reduce such errors, LLMs must be able to appropriately identify the type of research data described in a README file, and improving prompts to enable such distinctions remains future work.

5. Conclusion and Future Work

In this study, we investigated the feasibility of generating research data metadata from their accompanying README files. The results of the analysis indicated that README files tend to contain values that corresponded to metadata fields for research data and could serve as valuable resources for metadata generation. Then, we performed an experiment using LLMs, and the results demonstrated that the LLMs could extract metadata-related information from the README files with high performance.

In the analysis and experiment, we assumed that each README file described a dataset. However, in practice, the type of research data described in a README file is not always explicit. Thus, in the future, we plan to address the identification of the type of research data described in a README file.

6. Acknowledgements

This research was partially supported by the Grant-in-Aid for Scientific Research (B) (No. 25K03418) of JSPS and “Developing a Research Data Ecosystem for the Promotion of Data-Driven Science” of MEXT.

7. Bibliographical References

Zaid Alyafeai, Maged S. Al-shaibani, and Bernard Ghanem. 2025. MOLE: Metadata Extraction and Validation in Scientific Papers Using LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12236–12264.

DataCite Metadata Working Group. 2024. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and Other Research Outputs. Version 4.6.

Genet Asefa Gesese, Zongxiong Chen, Oussama Zoubia, and others. 2025. A Survey on Metadata for Machine Learning Models and Datasets: Standards, Practices, and Harmonization Challenges. In *Proceedings of 5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment (Sci-K 2025)*, volume 4065, pages 57–71.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Laura Koesten, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. 2020. Dataset Reuse: Toward Translating Principles to Practice. *Patterns*, 1(8).

Prince Kumar, Srikanth Tamilselvam, and Dinesh Garg. 2024. Read between the Lines - Functionality Extraction from READMEs. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3977–3990.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, and others. 2023. Efficient Memory Management for Large Language Model Serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023)*, page 611–626.

Mistral AI Team. 2024. Un Ministral, des Ministraux.

Open AI. 2025. Introducing GPT-5.

G7 OSWG. 2023. Annex 1: G7 Open Science Working Group (OSWG).

Anthony Cintron Roman, Kevin Xu, Arfon Smith, and others. 2023. Open Data on GitHub: Unlocking the Potential of AI. *arXiv preprint arXiv:2306.06191*.

David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 2021. SoMeSci- A 5 Star Open Data Gold Standard Knowledge Graph of Software Mentions in Scientific Articles. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (CIKM 2021)*, pages 4574–4583.

Yuxuan Shu and Vasileios Lampos. 2025. Sonnet: Spectral Operator Neural Network for Multivariable Time Series Forecasting. *arXiv preprint arXiv:2505.15312*.

Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 2023. Empowering Knowledge Discovery from Scientific Literature: A Novel Approach to Research Artifact Analysis. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 37–53.

Aarne Talman, Antti Suni, Hande Celikkanat, and others. 2019. Predicting Prosodic Prominence from Text with Pre-trained Contextualized Word Representations. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa 2019)*, pages 281–290.

UNESCO. 2021. UNESCO Recommendation on Open Science.

Carlos Utrilla Guerrero, Oscar Corcho, and Daniel Garijo. 2024. Automated Extraction of Research Software Installation Instructions from README Files: An Initial Analysis. In *Proceedings of the Natural Scientific Language Processing and Research Knowledge Graphs (NSLP 2024)*, pages 114–133.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2024. Capabilities and Challenges of LLMs in Metadata Extraction from Scholarly Papers. In *Proceedings of the 26th International Conference on Asia-Pacific Digital Libraries (ICADL 2024)*, pages 280–287.

Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 2025. Metadata Generation for Research Data from URL Citation Contexts in Scholarly Papers: Task Definition and Dataset Construction. In *Proceedings of the Third Workshop for Artificial Intelligence for Scientific Publications (WASP 2025)*, pages 72–79.

Mark Wilkinson, Dumontier Michel, IJsbrand Aalbersberg, and others. 2016. The FAIR Guiding Principles for Scientific Data Management and Stewardship. *Sci Data*, 3.

An Yang, Anfeng Li, Baosong Yang, and others. 2025. Qwen3 Technical Report. *arXiv preprint arXiv:2505.09388*.

8. Language Resource References

Papers With Code. 2019–25. *Papers With Code Datasets*. PID https://huggingface.co/pwc-archive/datasets. Originally obtained via the Papers With Code GitHub repository. The dataset is not currently directly distributed from the repository (Last Accessed: 2025-08-08).