研究データのメタデータ生成における README ファイルの利用

関戸康太郎 [1]　渡邉優 [2]　伊藤滉一朗 [2]　松原茂樹 [2,3]
[1] 名古屋大学情報学部　[2] 名古屋大学大学院情報学研究科
[3] 名古屋大学情報基盤センター
{sekido.kotaro.h0,watanabe.yu.x3}@s.mail.nagoya-u.ac.jp
{ito.koichiro.z5,matsubara.shigeki.z8}@f.mail.nagoya-u.ac.jp

概要

コードリポジトリは，従来は主にソフトウェア開発で利用されてきたが，近年では研究データリポジトリとしても利用されている．コードリポジトリで公開された研究データには README が付加されていることが多い一方で，メタデータが整備されていないことがある．そこで本論文では，研究データのメタデータを README を用いて生成することの実現可能性を検証する．まず，README におけるメタデータ生成に有用な情報の出現傾向を分析する．続いて，LLM を用いた抽出実験を行い，その抽出性能を評価する．

1　はじめに

コードリポジトリとは，GitHub[1)]などに代表される，主にソフトウェア開発で利用されてきたリポジトリである．コードリポジトリには，ソフトウェアのコードに加え，その利用方法などが記述されたドキュメントである README が付加されている．近年では，研究データの公開先としても利用される事例が増えており，コードリポジトリは研究データリポジトリとしての役割も担いつつある [1]．そこでは，研究データの概要や利用方法などの説明のために README が掲載されることが多い．

研究データが円滑に流通するには，メタデータが整備されていることが重要である [2]．メタデータとは，表 1 に示すように，体系化された項目とその値で表現されるデータであり，それらがメタデータリポジトリに登録されることで，研究データの検索性が高まる．しかし，コードリポジトリに掲載された研究データにはメタデータが整備されていないことも多く，公開されていても，その存在が十分に認知されにくい状況が生じている．

表 1　Helsinki Prosody Corpus [3] のメタデータ（一部）

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

そこで本論文では，研究データのメタデータを README を用いて生成することの実現可能性を検証する．README を用いてメタデータが生成されれば，それをメタデータリポジトリに登録することにより，コードリポジトリで公開された研究データの検索性が高まることが期待できる．

本研究ではまず，研究データのメタデータ生成に有用な情報を README がどの程度含んでいるかを GitHub を対象に分析した．その結果，メタデータ項目によって差はあるものの，README がメタデータ生成のための情報源として利用できることを確認した．次に，大規模言語モデル（LLM）を用いた抽出実験を行い，その抽出性能を評価したところ，メタデータ生成に有用な情報を README から高水準で抽出可能であることを確認した．

本論文の構成は以下の通りである．2 章では，研究データのメタデータと README について，関連研究を交えて説明する．続く 3 章では，研究データのメタデータ生成に有用な情報を README がどの程度含んでいるかを分析する．4 章では，抽出実験について報告する．最後に，5 章で本論文をまとめる．

1）https://github.co.jp/

2　研究データのメタデータ生成

2.1　研究データの流通とメタデータ

近年，オープンサイエンスの世界的な潮流にともない，研究データの公開及び流通の重要性が高まっている [4, 5]．研究データの流通を促進するためには，研究データのメタデータを整備することが重要である [2]．表 1 に，研究データのメタデータの例として，発話の抑揚を予測するためのデータセットである Helsinki Prosody Corpus [3] のメタデータの一部を示す．研究データのメタデータには，その名称（Title），提案年（PublicationYear），作成者（Creator），主題やキーワード（Subject），言語（Language）など，その研究データを特徴付ける情報が含まれる．

2.2　README を用いたメタデータ生成

研究データの README は，その利活用，すなわち，人間による研究データの理解と再利用を容易にするために作成される文書である．図 1 に，研究データの README の例として，Helsinki Prosody Corpus の README[2] の一部を示す．研究データの README には，研究データに関する説明が記載されており，メタデータ生成に有用な情報を含む可能性がある．例えば，図 1 に示した README には，研究データの名称（Title）や提案年（PublicationYear）に関する情報など，表 1 に示したメタデータに対応する情報が含まれている．コードリポジトリで公開された研究データの検索性を高める方策として，README を用いてメタデータを生成し，メタデータリポジトリに登録することが挙げられる．

2.3　関連研究

これまでに，研究データへの言及を含む文書から，研究データに関する情報の抽出が試みられている．その多くは，ソフトウェアやデータセットを対象に，学術論文からそれらのメタデータ生成に有用な情報を抽出している [6, 7, 8, 9]．

一方で，README から研究データに関する情報の獲得を目指した研究もいくつか存在する [10, 11]．これらは，ソフトウェアを対象に，その機能や利用方法などに関する言及部分の抽出を試みている．しかし，研究データのメタデータ生成に README を利用することは検討されていない．

2）https://github.com/Helsinki-NLP/prosody

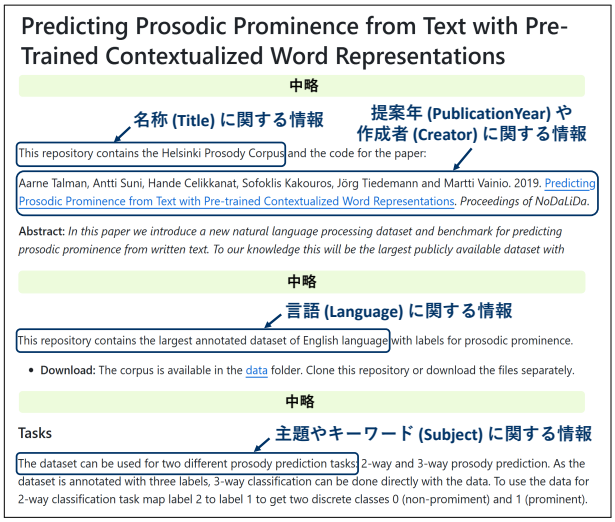

図 1　Helsinki Prosody Corpus [3] の README ファイル（一部）

3　分析

本研究ではまず，README がメタデータ生成のための情報源となりうるかを評価する．そのために，README におけるメタデータ生成に有用な情報の出現傾向を分析する．また近年では，コードリポジトリにデータセットが登録される事例が増加しているものの [1, 12]，ソフトウェアなどに比べ，これらのデータセットに着目した研究は限定的である．そこで本分析では，対象とする研究データの種別をデータセットとする．

3.1　分析データ

分析には，ある研究データに対するメタデータと README のペアで構成されるデータが必要となる．本研究では，Papers With Code Datasets [13]（以降，PWCD と表記）を利用して，分析データを作成した．PWCD には，データセットのメタデータが登録されており，データセットの公開先のリポジトリを参照する URL も含まれている．本研究では，公開先が GitHub である URL を対象に，GitHub API を用いて README を取得した[3]．取得したメタデータと README のペア 2,737 件のうち，README の取得元であるリポジトリの作成年が 2023 年以前である 2,464 件を分析データとした．

取得したデータに含まれるメタデータ項目の

3）複数の README が登録されているリポジトリについては，どの README がデータセットに対応するのかが自明ではなく，本分析の対象外とした．また，PWCD のメタデータは英語で記述されているため，README も英語で記述されているものに限定して取得した．

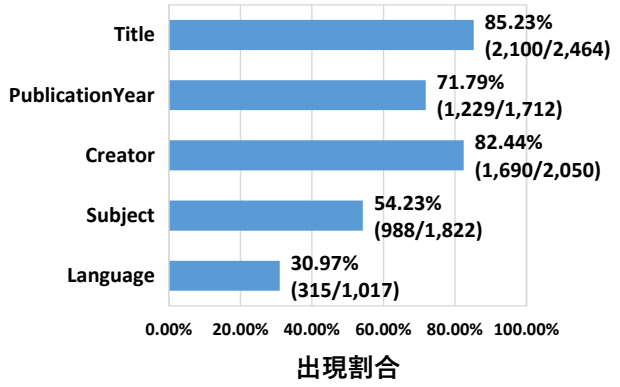

図 2　分析結果

うち，研究データのメタデータの国際標準である DataCite Metadata Schema [14] の項目に対応するものを確認した．その結果，表 1 で示した 5 つの項目について対応関係が認められたため，これらのメタデータ項目を分析対象とした．

3.2　分析方法

README におけるメタデータ項目の値の出現判定の方法は以下の通りである．まず，メタデータと README に対して，表記揺れの影響を抑制するための文字列処理を行った[4]．次に，メタデータ項目ごとに，その値が README に出現するか否かを判定した．PWCD のメタデータでは Creater，Subject，Language の値はリスト形式で表現されている．値がリスト形式の場合は，リストの要素が 1 つでも README に含まれれば，その値が README に出現したと判定した[5]．

3.3　分析結果

メタデータ項目ごとの出現割合を図 2 に示す．出現割合が高かった項目は Title と Creator であり，いずれも 80%を超えていた．最も出現割合が低かった Language でも 30%に達していた．分析対象である 5 つの項目の出現割合のマクロ平均は 64.93%であり，項目によって出現割合に差はあるものの，README がメタデータ生成のための情報源として利用できることを確認した．

4　実験

研究データのメタデータ生成に有用な情報を README から抽出することの実現可能性を検証するために，LLM を用いた抽出実験を行った．なお 3 章同様，本実験においても，データセットの README を対象とする．

4.1　実験データ

3 章で使用した分析データを訓練，開発用に分割し，分析に使用していない計 273 件のデータをテスト用に割り当てた．分割は訓練，開発，テスト用でおおよそ 8:1:1 の比率になるように行った．本実験では，開発データをプロンプト設計のために用い，テストデータを評価のために用いた．

4.2　実装

実験で使用した LLM について説明する．オープンモデルとしては，Llama-3.1-8B-Instruct [15], Ministral-8B-Instruct-2410 [16], Qwen3-8B [17], クローズドモデルとしては，GPT5 [18] を用いた．オープンモデルと GPT5 による抽出を行うために，vLLM [6] と OpenAI API [7] をそれぞれ利用した．なお，vLLM については temperature を 0.0 に設定した．すべてのモデルに Structured Outputs を適用し，表 1 に示す 5 つの項目からなるメタデータを JSON 形式で出力するよう指示した．プロンプトでは，データセットのメタデータ項目の値を抽出するように指示し，抽出する値が特定できない場合は null を出力するよう指示した[8]．

4.3　評価方法

メタデータ項目ごとに，LLM の性能を評価した．

評価指標　評価指標には再現率，適合率，F 値を用いた．それぞれの指標は以下の方法で算出した．

- 再現率 : README に出現したメタデータ項目の値のうち，正しく抽出できたものの割合
- 適合率 : LLM が出力した null でない値のうち，正しいものの割合
- F 値 : 再現率と適合率の調和平均

正誤判定方法　抽出された値の正誤を，正解文字列と LLM の出力文字列の一致により判定した．文字列一致の判定方法は Title, Subject は部分一致，PublicationYear, Creator, Language は完全一致とした[9]．また，メタデータ項目の値がリスト形式の

4）　文字列処理の詳細は付録 A を参照されたい．
5）　PWCD のメタデータの中には，値が未入力の項目が存在する．値が未入力であった場合は，README におけるメタデータ項目の値の出現割合を算出する際の対象外とした．
6）　https://github.com/vllm-project/vllm
7）　https://openai.com/ja-JP/api/
8）　実験に使用したプロンプトは，付録 B を参照されたい．
9）　一致判定を行う際，正解文字列と出力文字列の両方に，3 章の分析と同じ文字列処理を項目ごとに行った．

表 2　実験結果（再現率と適合率）．太字は各項目で最も高い数値．

	再現率				適合率			
	Llama	Ministral	Qwen	GPT	Llama	Ministral	Qwen	GPT
Title	82.68 (191/231)	79.65 (184/231)	84.42 (195/231)	**93.07 (215/231)**	72.43 (197/272)	71.43 (190/266)	77.44 (206/266)	**83.71 (221/264)**
PublicationYear	93.91 (185/197)	93.91 (185/197)	**94.92 (187/197)**	94.42 (186/197)	85.34 (198/232)	**89.72 (192/214)**	84.17 (202/240)	86.09 (198/230)
Creator	**98.42 (187/190)**	96.84 (184/190)	95.26 (181/190)	96.32 (183/190)	93.97 (187/199)	90.64 (184/203)	87.44 (181/207)	**97.34 (183/188)**
Subject	81.13 (86/106)	79.25 (84/106)	73.58 (78/106)	**92.45 (98/106)**	54.49 (91/167)	51.91 (95/183)	46.77 (87/186)	**67.72 (128/189)**
Language	74.19 (23/31)	**90.32 (28/31)**	**90.32 (28/31)**	77.42 (24/31)	87.91 (80/91)	93.33 (126/135)	92.62 (138/149)	**94.25 (82/87)**

表 3　実験結果（F 値）．太字は各項目で最も高い数値．

	Llama	Ministral	Qwen	GPT
Title	77.22	75.32	80.78	**88.14**
PublicationYear	89.42	**91.77**	89.22	90.06
Creator	96.14	93.64	91.18	**96.83**
Subject	65.19	62.73	57.19	**78.18**
Language	80.47	**91.80**	91.46	85.01
マクロ平均	81.69	83.05	81.97	**87.64**

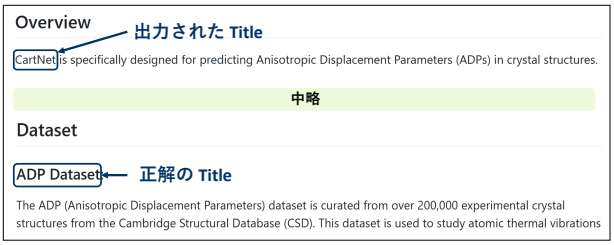

図 3　抽出に失敗した ADP Dataset [19] の README（一部）

場合は，リストの要素が 1 つでも一致すれば正解とした．テストデータ内のメタデータ項目の値が未入力の場合は判定を行わず，評価の対象外とした．

4.4　実験結果

再現率と適合率の結果を表 2 に示す．再現率については，いずれのメタデータ項目においても最も高いモデルでスコアが 90%を上回っており，LLM が README に含まれるメタデータ項目の値を高い網羅性で抽出できることを確認した．適合率については，Subject のスコアが相対的には低かったものの，その他のメタデータ項目においては最も高いモデルのスコアが 80%を上回っており，高い正確性でメタデータ項目の値を抽出できていることを確認した．

次に，F 値の結果を表 3 に示す．いずれのモデルも，F 値のマクロ平均が 80%を上回っており，その中でも GPT5 の抽出性能が最も高かった．これらのことから，LLM を用いることで，研究データのメタデータ生成に有用な情報を README から高水準で抽出可能であることを確認した．

4.5　エラー分析

LLM は高い抽出性能を示したものの，誤った抽出も確認された．本実験では，README 内にデータセットとは別の種別の研究データが出現する場合において，その研究データに関する情報を抽出してしまう誤りを確認した．その具体例として，原子の熱振動解析のためのデータセットである ADP Dataset [19] の README[10] の一部を図 3 に示す．この例では，正解の Title は ADP Dataset であるにもかかわらず，ADP Dataset で学習されたモデルである CartNet が誤って Title として抽出されている．このような誤りを抑制するためには，README 内に記述された研究データの種別を LLM が適切に判定できる必要がある．そのためのプロンプトの改良は今後の課題の 1 つである．

5　おわりに

本論文では，研究データのメタデータを README を用いて生成することの実現可能性を検証した．README を分析し，README には研究データのメタデータ項目の値が含まれやすく，メタデータ生成のための情報源として利用できることを確認した．また，LLM による抽出実験を行い，研究データのメタデータ生成に有用な情報を README から高水準で抽出可能であることを確認した．

本実験では，README がデータセットのものであることを前提としたが，一般的には，README の対象は自明ではない．今後は，README の記述対象の識別にも取り組みたい．

10）https://github.com/imatge-upc/CartNet

謝辞

本研究は，JSPS 科研費 JP25K03418 の助成，文部科学省「AI 等の活用を推進する研究データエコシステム構築事業」の支援を受けたものです．

参考文献

[1] Laura Koesten, Pavlos Vougiouklis, Elena Simperl, and Paul Groth. Dataset Reuse: Toward Translating Principles to Practice. **Patterns**, Vol. 1, No. 8, 2020.

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

[3] Aarne Talman, Antti Suni, Hande Celikkanat, and others. Predicting Prosodic Prominence from Text with Pretrained Contextualized Word Representations. In **Proceedings of the 22nd Nordic Conference on Computational Linguistics**, pp. 281–290, 2019.

[4] UNESCO. Open science. https://www.unesco.org/en/open-science (Last Accessed 2025/12/08).

[5] Annex 1: G7 Open Science Working Group (OSWG). https://www8.cao.go.jp/cstp/kokusaiteki/g7_2023/annex1_os.pdf (Last Accessed 2025/12/08).

[6] David Schindler, Felix Bensmann, Stefan Dietze, and Frank Krüger. 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)**, pp. 4574–4583, 2021.

[7] Petros Stavropoulos, Ioannis Lyris, Natalia Manola, Ioanna Grypari, and Haris Papageorgiou. 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)**, pp. 37–53, 2023.

[8] Yu Watanabe, Koichiro Ito, and Shigeki Matsubara. 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)**, pp. 280–287, 2024.

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

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

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

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

[13] Papers With Code Datasets. https://github.com/paperswithcode/paperswithcode-data (Last accessed 2025/08/08).

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

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

[16] Mistral AI Team. Un Ministral, des Ministraux. https://mistral.ai/news/ministraux (Last accessed 2025/12/07).

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

[18] Open AI. Introducing GPT-5. https://openai.com/ja-JP/index/introducing-gpt-5/ (Last accessed 2025/12/07).

[19] Àlex Solé, Albert Mosella-Montoro, Joan Cardona, and others. A Cartesian Encoding Graph Neural Network for Crystal Structure Property Prediction: Application to Thermal Ellipsoid Estimation. **Digital Discovery**, Vol. 4, pp. 694–710, 2025.

A　文字列処理方法

3 章の分析で出現判定の際に行った，メタデータと README への文字列処理の内容を表 4 に示す．標準化には，小文字化と記号類の除去が該当し，略語化には，文字列の大文字を連結する処理が該当する．以下に，標準化と略語化の例を示す．

- 標準化の例
  - 標準化前：Brno-Urban-Dataset
  - 標準化後：brno urban dataset
- 略語化の例
  - 略語化前：Brno-Urban-Dataset
  - 略語化後：BUD

表 4　文字列処理の対応

項目	メタデータ	README
Title	略語化	処理なし
	標準化	標準化
PublicationYear	処理なし	処理なし
Creator Subject Language	リストの要素をそれぞれ標準化	標準化

B　プロンプト

抽出実験で用いたプロンプトを図 4 に示す．このプロンプトは，開発データを用いた予備実験を踏まえて作成されたものである．なお，プロンプト内のメタデータの項目名は，抽出する値の定義を明確にするための表現を用いている．

```
# system prompt
You are a metadata extraction model.
You will be provided 'README Text'.
Your task is to extract dataset metadata from the provided
'README Text'.
If a value is **not explicitly stated**, you may **carefully
infer** it based on clear context.
You must return a valid JSON object that strictly follows this
schema:
{
"name": string or null,
"introduced_year": integer or null,
"creators": list of strings or null,
"tasks": list of strings or null,
"natural_languages": list of strings or null,
}
If neither explicit information nor a clear inference is possi-
ble, set the value to null.

# user prompt
Extract the dataset metadata from the README Text.
README Text:
{readme_text}
```

図 4　抽出実験で用いたプロンプト