研究データのオープンアクセスを加速化する：生成 AI を用いたメタデータ生成と機関リポジトリへの収載

渡邉 優 [1)], 茂木 光志 [1)], 松原 茂樹 [2),1)]

1) 名古屋大学大学院情報学研究科
2) 名古屋大学情報基盤センター

watanabe.yu.x3@s.mail.nagoya-u.ac.jp

Accelerating Open Access to Research Data: Metadata Generation Using Generative AI and Its Deposition in Institutional Repositories

Yu Watanabe[1)], Koshi Motegi[1)], Shigeki Matsubara[2),1)]

1) Graduate School of Informatics, Nagoya University
2) Information Technology Center, Nagoya University

概要

本論文では，大学における研究データのオープンアクセスの加速化を目指し，研究データのメタデータ生成とその機関リポジトリへの収載の新たな方式を提案する．大学における研究データ公開は，メタデータ作成の負担が小さくなく，十分に進んでいないのが現状である．本方式では，学術論文テキストおよび公開プラットフォーム内の情報を活用し，大規模言語モデル（LLM）によりメタデータを自動生成し，それをリポジトリソフトウェアに搭載する．論文テキストデータを用いた実験では，同定された 16,426 件の研究データの 64.8% に相当する 10,640 件についてメタデータが生成され，データリポジトリの効率的作成の実現性が示された．

1 はじめに

オープンサイエンスとは, 研究データを公開し, その利活用を促進する取り組みである．オープンサイエンスの到達度を示す基準として，FAIR 原則がある [1]. これは, *Findable*, *Accessible*, *Interoperable*, *Reusable* という要素の総称である．このうち，最も基本的な原則は *Findable* であり，研究データが見つけられるためには,

- 一意な識別子（ID）が付与されている,
- メタデータが十分に記されている

ことが求められている．

世界的なオープンサイエンスの潮流のもと [2]，国内の学術機関でも研究データの公開に向けた整備が進みつつある．例えば，これまでに策定された大学の研究データポリシーの多くは,

- 研究者は研究データを可能な限り公開すること
- 大学は研究者による研究データ公開を支援すること

を謳っている（例えば，[3]）．このポリシーを遂行するための大学の施策として，研究データの機関リポジトリへの収載が推奨されている．機関リポジトリへの研究データ収載において実績を高め，大学の価値向上につなげることが課題となっている．

国や大学の施策や指針が整備される一方，研究者が研究データを大学の機関リポジトリに掲載する動きは十分に進んでいない．その要因として,

- 研究データのオープン化に関する慣習や指針は，研究分野で定まっていることが多く，それを転換することは必ずしも容易ではない
- 研究データを機関リポジトリに登録するためには，メタデータを作成する必要があり，それを担う研究者ならびに支援者の負担は小さくない

ことが挙げられる．今後，大学における研究データのオープン化を進展させるには，上述の要因を解消する新たな仕組みの導入が求められる．

本論文では，大学における研究データのオープンアクセスの加速化に向け，研究データのメタデータ生成

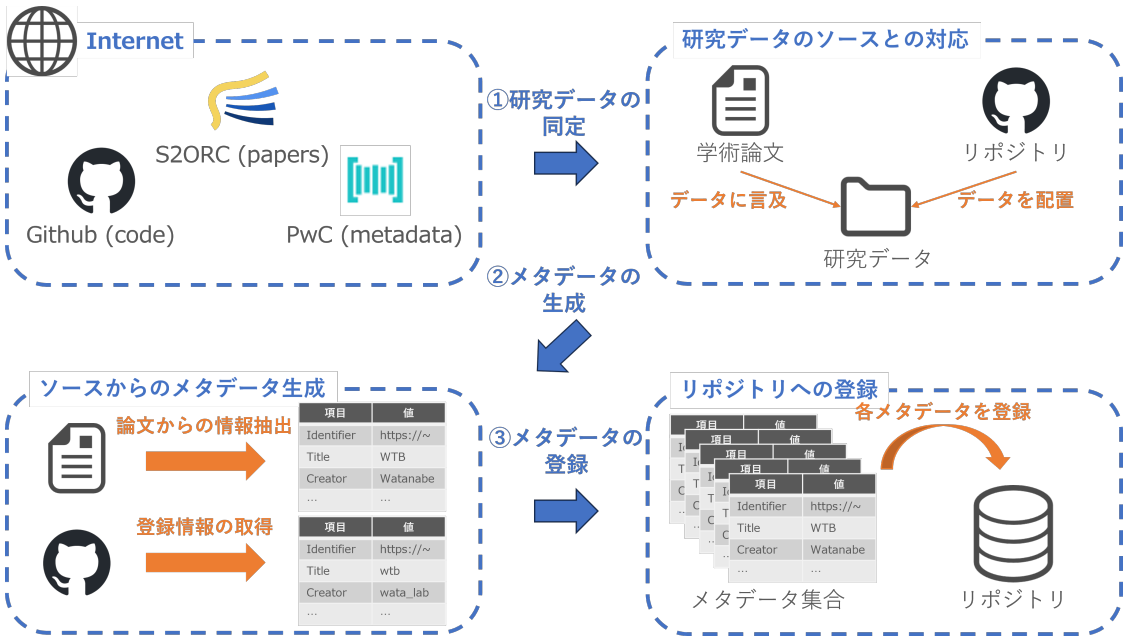

図 1　研究データのメタデータリポジトリの自動構築

とその機関リポジトリへの収載の新たな方式を提案する．これまで主に手作業で行われていた研究データの機関リポジトリへの収載が自動化されることにより，研究者や支援者の労力が軽減されることが期待される．

本方式では，メタデータ生成のための情報源として以下を活用する．

- 学術論文テキストに記載された研究データに関する情報
- データセットやコードを共有する研究データ公開プラットフォームに格納された情報

情報源からのメタデータの自動生成に大規模言語モデル（LLM）を活用する．生成されたメタデータはリポジトリソフトウェアに機械的に搭載できるため，研究データのメタデータリポジトリを効率的に構築できる．

本開発では, 学術論文データベースとして S2ORC[4] を, 研究データ公開プラットフォームとして GitHub[*1] を，また，リポジトリソフトウェアとして WEKO3[5] を，それぞれ使用している．実験では，学術論文を用いて同定した 16,426 件の公開された研究データに対して，その 64.8% に相当する 10,640 件についてメタデータを生成できた．それらを WEKO3 に一括登録することで，LLM を用いたメタデータの自動生成とその機関リポジトリへの収載の実現性を確認した．

[*1] https://github.com/

本論文の構成は以下の通りである．2 章では，研究データのオープン化の現状について論じる．3 章では，研究データのメタデータをリポジトリに収載する方式について概説する．4 章ではメタデータの自動生成について，5 章ではメタデータの登録について述べる．

2　研究データのオープンアクセスの現状

内閣府統合イノベーション戦略推進会議による「公的資金による研究データの管理・利活用に関する基本的な考え方」[6] では，オープンサイエンス推進への取り組みに関する基本方針が示されている．研究者の責務として，

- メタデータを付与し研究データ基盤システムで検索できるように登録すること
- オープン・アンド・クローズ戦略に従い，研究データの公開・共有を行うこと

また，大学等の機関の責務として，

- 研究データを機関リポジトリに収載すること
- 研究データへのメタデータ付与を推進すること

が示されている．

オープンサイエンスに関する国際的潮流や我が国におけるオープンアクセスの施策の影響もあり，大学によっては機関リポジトリに研究データの収載を進めるための仕組みを整えつつあるものの，その動きは十分に進んでいるとは言い難い．この理由として，

表 1　PwC Datasets に登録された研究データの例

項目	値
name	WikiSQL
homepage	https://github.com/salesforce/WikiSQL
description	**WikiSQL** consists of a corpus 87,726 hand-annotated SQL query ...
paper.title	Seq2SQL: Generating Structured Queries from Natural Language using Reinforcement Learning
modalities	Texts

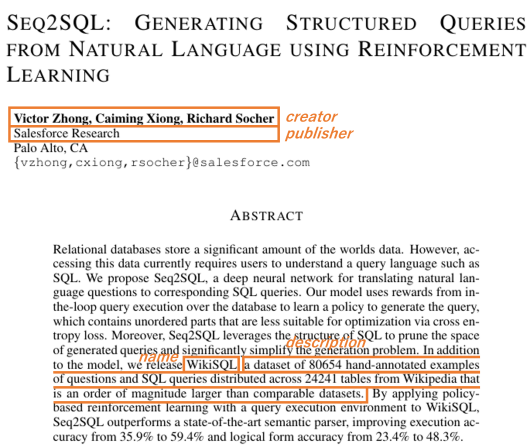

図 2　WikiSQL の論文（[7] より引用）

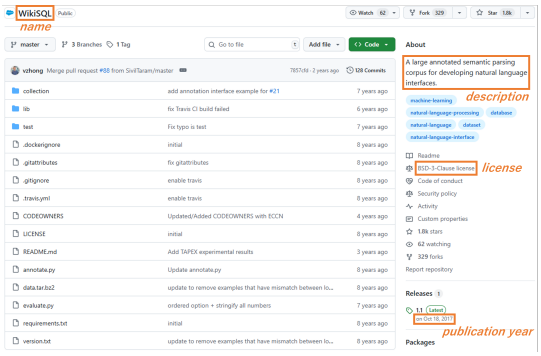

図 3　WikiSQL の GitHub[*3]

- 何を，いつ，どこに掲載するかなど研究データのオープン化に関する慣習や方針は，研究分野や研究グループごとに定まっていることが多く，それらを急速に転換することは容易ではない，
- 機関リポジトリに新たに研究データを登録するためには，共通化仕様のメタデータを別途作成する必要があり，それを担う研究者ならびに支援者の労力は小さくない，

ことが挙げられる．すなわち，研究データを公開する動きはあるものの，それを機関リポジトリを介して実施する動きに至っていないのが現状である．これを打破する方策として，公開されている研究データに対してメタデータを生成し，それを機関リポジトリに格納することが考えられる．

3　提案方式の概要

本方式では，インターネット上に公開されている研究データに対してメタデータを生成し，それをリポジトリに収載することでデータリポジトリを構築する．これは以下の手順で実現される．

1. 情報公開された研究データのうち，情報源に紐づけが可能なものを同定する，
2. 紐づけられた情報源から研究データに関する情報を獲得し，メタデータを生成する．
3. 生成されたメタデータをリポジトリに搭載する．

図 1 に，本方式に基づくデータリポジトリ構築の流れを示す．

手順 1 では，学術論文リポジトリ，コードリポジトリ，メタデータリポジトリ等を利用して，公開された研究データを同定し，学術論文や研究データが配置されているリポジトリを紐づける．**手順 2** では，学術論文テキストからの情報抽出，及び，リポジトリの登録情報の取得により，研究データのメタデータを自動生成する．**手順 3** では，メタデータリポジトリのフレームワークに対して，生成した各メタデータを登録する．

これらの手順を実行することにより，大量の研究データのメタデータを効率的に作成でき，これを公開することでオープンサイエンスへの貢献が期待できる．

4　研究データのメタデータ生成

本章では，研究データのメタデータリポジトリの構築手順のうち，1 及び 2 について述べる．

4.1　情報源に紐づけ可能な研究データの同定

単にインターネット上に公開されているだけの研究データに対してメタデータを生成することは現実的で

[*3] https://github.com/salesforce/WikiSQL
[*4] 本研究では，Contributor を Publisher と同一であると見なした．

表 2　DataCite [8] のメタデータ項目（一部抜粋）

項目	定義	種別	ソース
Identifier	The Identifier is a unique string that identifies a resource.	必須	論文
Creator	The main researchers involved in producing the data, or the authors of the publication, in priority order.		両方
Title	A name or title by which a resource is known.		両方
Publisher	The name of the entity that holds, archives, publishes, prints, distributes, releases, issues, or produces the resource.		両方
PublicationYear	The year when the data was or will be made publicly available.		GitHub
ResourceType	A description of the resource.		論文
Subject	Subject, keyword, classification code, or key phrase describing the resource.	推奨	論文
Contributor [*4]	The institution or person responsible for collecting, managing, distributing, or otherwise contributing to the development of the resource.		論文
Date	Different dates relevant to the work.		GitHub
Description	A description of the resource.		両方
Rights	Any rights information for this resource.	任意	GitHub

ない．本方式では，関連する情報源が存在しそれらと紐づけ可能な研究データを同定する．そのような情報源として，

- 既存のメタデータリポジトリ内のメタデータ
- 学術論文において研究データを参照するテキスト

を使用することが考えられる．

前者のメタデータリポジトリは，様々な分野や目的で運用されており，公開された研究データに対するメタデータを収載している．これらは，学術機関リポジトリのものとは仕様が異なるものの，新たなメタデータの生成に利用できる．後者については，学術論文では作成あるいは使用した研究データについて言及されるため，そのテキストはメタデータ生成の情報源として利用できる．

本開発では，研究データの同定のために以下に示す 2 つの方法を採用した．

4.1.1　論文と研究データの対応データの利用

PwC Datasets[*5]を使用して，論文を情報源として利用可能な研究データを同定した．これは，機械学習で使用されるデータセットに対してユーザが情報を付加することが可能なプラットフォーム “Papers with Code” のバックエンドで使用されている．

表 1 に，本データセットに格納されている研究データの情報の例を示す．格納されている情報のうち，研究データのホームページ URL（表 1 の homepage）を研究データの識別子として同定した．研究データと論文の紐づけには，論文タイトル（paper.title）を利用し，タイトルの一致により S2ORC の論文テキストと紐づけた．また，GitHub の紐づけにはホームページの URL（homepage）を使用し，その遷移先 GitHub リポジトリと紐づけた．図 2 と図 3 に，WikiSQL という研究データに紐づいた学術論文と GitHub リポジトリをそれぞれ示す．

4.1.2　学術論文コーパスの利用

論文コーパス S2ORC を使用して研究データを同定した．S2ORC は，数百万件の学術論文のメタデータと抄録を収集した大規模なコーパスである．本コーパスの論文テキストは，論文データを PDF Parser により解析し，構造化情報を付与することで作成されている．本開発では，この中でフルテキストが格納された論文に対して研究データの同定を行った．

研究データの同定は，学術論文テキストに対して研究データを参照する URL を検出することにより行った．学術論文テキストから正規表現を用いて URL を検出し[*6]，これを研究データの識別子とみなした．このとき，抽出元となった学術論文，および，検出された URL の遷移先 GitHub リポジトリをそれぞれ研究データに紐づけた．

[*5] https://github.com/paperswithcode/paperswithcode-data

[*6] 正規表現は，
`r'\b(?:is|are|was|were|be|been|being)?\s*(?:available|accessible|downloadable)?\s*(?:at|from)\s+((?:https?|ftp)://(?:www\.)?github\.com/[^\s"'"()<>,;]*)'` を用いた．

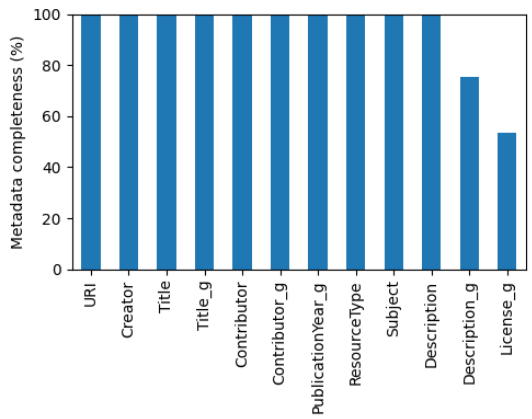

図 4　各研究データに対するメタデータの充足率．“_g”は GitHub から取得した項目を示す．

4.2　研究データに関する情報の獲得

研究データのメタデータを生成し，機関リポジトリに登録するには，共通化仕様のスキーマに従う必要がある．本開発では，国際標準の汎用性を備えている DataCite メタデータスキーマ [8] を採用した．表 2 に，本論文で生成の対象とするメタデータ項目を示す．表 2 の各項目を生成するため，学術論文あるいは GitHub を用いて情報を獲得する．

学術論文を用いた研究データ情報の獲得では，生成 AI を用いた情報抽出により行った．情報抽出のための生成 AI として，OpenAI API を介して，GPT-4o-mini[*7]と GPT-4.1-nano[*8]を利用した．メタデータ生成実験の期間中に廉価版のモデルが登場したため，mini モデルから nano モデルへ切り替えている．なお，生成 AI に対してメタデータに該当する文字列を厳格に出力させるため，Structured Outputs 機能[*9]を利用した．抽出対象は，表 2 のソースが「両方」と「論文」である項目とした．

GitHub に登録された情報の取得には，PyGithub[*10]を利用した．PyGithub は Python から GitHub API にアクセスするライブラリである．このライブラリを用いて，表 2 のソースが「両方」と「GitHub」である項目を取得した．

[*7] https://platform.openai.com/docs/models/gpt-4o-mini
[*8] https://platform.openai.com/docs/models/gpt-4.1-nano
[*9] https://platform.openai.com/docs/guides/structured-outputs
[*10] https://github.com/PyGithub/PyGithub

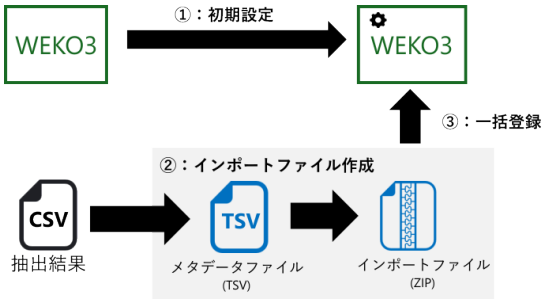

図 5　メタデータ登録の概略図

4.3　獲得された情報によるメタデータ項目の充足性

以下では，提案方式を実装し，メタデータ生成に適用した結果について述べる．

既存のメタデータリポジトリ PwC Datasets の利用，あるいは，論文コーパス S2ORC の一部の利用により，論文と GitHub の組でそれぞれ 3,943 件，12,483 件取得した[*11]．この中で，GitHub で公開された研究データのメタデータを 10,640 件生成できた．図 4 に，メタデータの生成の対象とした研究データの総数における各メタデータ項目の充足率を示す．各研究データに対してほとんどの項目のメタデータを自動で生成できていることを示している．その一方で，GitHub からの Description と License の取得率が低いことが分かった．これらの項目は GitHub リポジトリの作成者が任意で登録するものであり，登録されることが少ないことが起因している．以上のことから，研究データの URL を識別子として同定できれば，十分なメタデータを生成できるといえる．

5　メタデータのリポジトリへの収載

本章では，生成したメタデータのリポジトリへの収載について述べる．

本研究における登録の過程を図 5 に示す．リポジトリソフトウェアとして WEKO3 を使用する．WEKO3 は，国立情報学研究所 (NII) オープンサイエンス基盤研究センター (RCOS) で開発されており，オープンサイエンスリポジトリ推進協会 (JPCOAR) が共同で運用するクラウドサービス JAIRO Cloud で採用されている．

5.1　メタデータ収載の手順

メタデータの登録には，WEKO3 のアイテム一括登録（インポート）機能を利用する．一括登録のための操作は，以下の手順で実行できる．

[*11] 論文と GitHub が 1 対 1 対応する組のみを扱った．

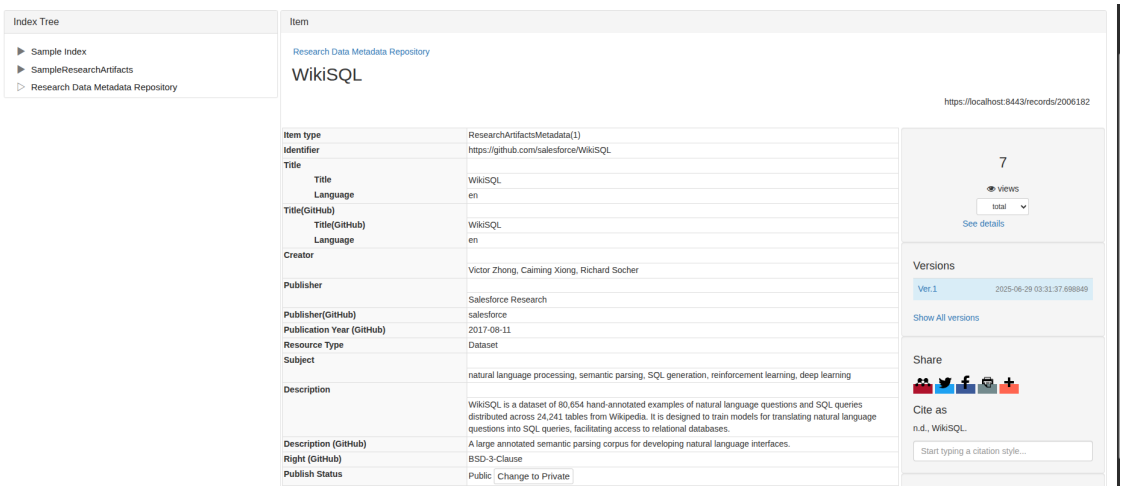

図 6　WEKO3 のメタデータの画面．(GitHub) はソースが GitHub であることを示す．

1. WEKO3 を初期設定する
2. インポートファイルを作成する
3. インポートファイルを用いて一括登録する

このうち，手順 2 と 3 はメタデータ登録の際に毎回必要となる．これらの操作を設定することで，リポジトリの運用を自動化できる．各手順の詳細を以下に記す．

手順 1　WEKO3 に一括登録するために以下の項目を設定する．

- アイテムタイプ：メタデータのデータ型
- インデックス：登録アイテムをまとめる単位
- ワークフロー：アイテム登録から公開までの操作列

本研究では，GitHub で公開されているソースコード[*12]を用いて WEKO3 を起動し，表 2 に示すデータ項目に対応するアイテムタイプを作成した．続いて，メタデータ登録先のインデックス，及び，そのワークフローを設定した．

手順 2　WEKO3 に一括登録するためのインポートファイルを作成するため，以下の手順を実行する．

1. 各研究データのメタデータを所定の形式で記述したメタデータファイル (TSV 形式) を作成する
2. メタデータファイルを含むファイル群を圧縮し，インポートファイル (ZIP 形式) を作成する

本研究では，研究データは登録せず，4.2 節で生成したメタデータのみを登録するため，メタデータファイルのみを所定の位置に配置したフォルダを圧縮し，インポートファイルを作成する．なお，これらの操作は，作成したプログラムを用いて自動化している．

手順 3　作成したインポートファイルを用いて WEKO3 に一括登録する．具体的には，WEKO3 のインポート機能の操作手順[*13]に従ってインポートファイルをアップロードする．なお，本研究では，Selenium[*14]を用いて操作を自動化している．

5.2　メタデータ登録の実行

図 6 に，登録されたメタデータに対する WEKO3 のスクリーンショットを示す．一連の実行を通して，リポジトリソフトウェア WEKO3 を用いることにより，生成された研究データのメタデータをリポジトリに一括登録できることを確認した．

6　まとめ

本論文では，大学における研究データのオープンアクセスの加速化を目指し，研究データのメタデータ生成とその機関リポジトリへの収載の新たな方式を提案した．本方式では，学術論文および公開プラットフォームに含まれる情報を活用し，LLM を用いてメタデータを自動生成し，それらをリポジトリソフトウェアに収載する．論文データベースを用いた実験では，同定された 16,426 件の研究データの 64.8% に相当する 10,640 件についてメタデータが生成され，データリポジトリの効率的作成の実現性が示された．

[*12] https://github.com/RCOSDP/weko

[*13] https://meatwiki.nii.ac.jp/confluence/spaces/JAIROCloudWEKO3/pages/63868511/アイテム一括登録（インポート）

[*14] https://www.selenium.dev/ja/

謝辞　本開発は，一部，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．日頃からご議論いただく本事業の関係の皆様，特に，「ルール・ガイドライン整備チーム」のメンバー諸氏に感謝致します．

参考文献

[1] Mark Wilkinson, Michel Dumontier, IJsbrand Aalbersberg, et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, Vol. 3, No. 160018, 2016.

[2] OECD. Making open science a reality. Technical Report 25, OECD Publishing, Paris, 2015.

[3] 名古屋大学. 学術データポリシー, 2020. `https://icts.nagoya-u.ac.jp/ja/datapolicy/`.

[4] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4969–4983, Online, July 2020. Association for Computational Linguistics.

[5] 国立情報学研究所 オープンサイエンス基盤研究センター. Weko3（公開基盤）. `https://rcos.nii.ac.jp/service/weko3/`, 2025. Accessed: 2025-09-26.

[6] 統合イノベーション戦略推進会議. 公的資金による研究データの管理・利活用に関する基本的な考え方. 内閣府, 2021.

[7] Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *CoRR*, Vol. abs/1709.00103, , 2017.

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