

第 25 回情報知識学フォーラム予稿

DataCite を利用した日本の大学における 研究データの公開状況についての分析

An Analysis of Current State of Research Data Publication at Japanese Universities

西岡 千文^{1*}

Chifumi NISHIOKA^{1*}

¹ 京都大学附属図書館

Kyoto University Library, Kyoto University

〒 606-8501 京都市左京区吉田本町

E-Mail: nishioka.chifumi.2c@kyoto-u.ac.jp

* 連絡先著者

オープンサイエンスの潮流の中、研究データの公開が推進されている。本稿では、日本の大学における研究データの公開状況について分析を行う。分析では DataCite の全レコードをデータセットとして使用した。全レコードのうち、日本の大学に所属している著者が含まれているレコードは 39,128 件あることがわかった。このうち半数以上が、高エネルギー物理学のリポジトリである HEPData に登録されたレコードであった。分析より、カウント方法によって大きく研究データ公開件数の結果が異なること、近年研究データ公開件数は増加傾向にあること等が明らかとなった。

In the era of Open Science, a lot of researchers committed to research data publication. This article analyzes the current state of research data publication at Japanese universities, using records of DataCite. We identify 39,128 records, in which one or more authors are affiliated with a Japanese university. More than half of the records are available on HEPData, a repository of high energy physics. We observe that the number of research data publications differ greatly depending on counting methods. In addition, the result shows that the number of research data publications has been increased in recent years.

キーワード: 研究データ, 学術情報流通, オープンサイエンス

Keywords: research data, scholarly communication, open science

1 はじめに

近年、オープンサイエンスの潮流の中で、政府機関や研究機関によって研究データの公開がより一層推進されている。日本における研究データの公開状況については、アンケート等で調査が行われてきた^[1]。本稿では、研究データに DOI の割り当てを行っている DataCite の全レコードのメタデータを使用することによって、日本の大学における研究データ公開状況の分析を試みる。

本稿の構成を以下に示す。続く章では分析の手法について述べ、3 章で分析の結果を報告する。4 章では、本稿をまとめるとともに、分析の課題、今後の方向性を示す。

2 手法

本章では分析に使用したデータセット (2.1 節)、日本の大学に所属する著者が含まれるレコードの抽出 (2.2 節)、分析の対象とするレコードのリソースタイプ (2.3 節)、本稿で採用する研究データ公開件数のカウント方法 (2.4 節) について述べる。

2.1 データセット

分析では DataCite の全レコードのメタデータ 19,606,708 件をデータセットとして使用した^[2]。DataCite は研究データを中心としたリソースに DOI を割り当てる機関であることから分析の目的と合致する。

2.2 著者の所属機関の抽出と同定

DataCite の各レコードには著者のリストがフィールド `creators` に、各著者の所属機関のリストがフィールド `affiliation` に格納されている。フィールド `affiliation` の各要素には著者の所属機関が文字列で入力されているが、表記揺れ（例：“Kyoto Univ.”, “Kyoto University”）がある。表記揺れを解決し、日本の機関に所属する著者が含まれているレコードを抽出するために、各レコードのメタデータから全著者の機関を抽出し、NISTEP 機関同定プログラム公開版と機関名辞書（2020 年 8 月 28 日、試用実験用に配布された版）^[3] で同定を行った。機関同定プログラムは、与えられた文字列に対して、対応する機関の機関名、ID、セクターなどを出力する。レコードに含まれている著者は延べ 200,011,615 名、そのうち所属機関の情報がない著者は延べ 101,440,687 名であった。著者の所属機関延べ 98,603,753 件について、機関同定プログラムで表記揺れの解決を行った。1 名の著者が複数の機関に所属するケースがあるため、著者の所属機関の延べ数は、著者の延べ数から所属機関の情報がない著者の延べ数を差し引いた数よりも大きい。

その結果、日本の機関に所属している著者が含まれるレコードは 64,953 件存在することがわかった。機関同定プログラムの結果では、機関のセクターが国の機関であったり、国立研究開発法人等である際に、海外の機関であるにも関わらず日本の機関が同定されてしまうケースが多々観察された。機関同定プログラムは日本の機関のみに対応しており、入力ファイルでは全ての文字列の国コードの欄には便宜的に `jpn` を入力していることから、このような誤りが発生した。よって、セクターが公立大学、国立高専、国立大学、国立短大、私立大学、私立短大のいずれかである機関に所属する著者が含まれているレコード 43,651 件を対象とする。

2.3 対象とするリソースタイプ

DataCite は研究データを主な対象として DOI を割り当てているが、レコードには論文、発表資料といったリソースも含まれている。論文、発表資料といったリソースは研究データではないので、調査から除外するべきである。DataCite のメタデータスキーマ 4.0 版以降ではフィールド `resourceTypeGeneral`

の入力が必須となっており、Dataset といったリソースの種別を表現する統制語彙が入力されている^[4]。研究データに相当する統制語彙については様々な解釈があるが、本稿ではリソースの種別が Collection, Dataset, Software のうちいずれかであるレコード 39,128 件を分析の対象とした。

2.4 研究データ公開件数カウント方法

本稿では以下の 2 要因に基づき、日本の大学における研究データ公開件数を 6 通りの方法で算出する。

- (A) カウント方法：整数カウント or 分数カウント or 第一著者
- (B) HEPData で公開されているレコード：含める or 除く

以下でそれぞれの要因と選択肢について述べる。

(A) カウント方法についてであるが、整数カウントでは 1 件のレコードの著者の所属機関に複数の国が出現していたとしても、いずれの国についてもレコードを 1 件とカウントする。よって日本の大学に所属する著者がレコードに含まれていれば、1 件としてカウントする。対して分数カウントでは、レコード 1 件を著者数、著者が所属する機関数、さらに同定される機関数（機関名が曖昧な際には、機関同定プログラムは 1 件の文字列に対して複数の機関を同定する）で按分する。論文の件数に基づき研究力等の評価を行う際には、各論文の全著者を考慮するケースの他に、筆頭著者や責任著者として執筆した論文のみをカウントするケースがある。本稿では研究データの公開は第一著者が主導して行っていると想定して、第一著者のみをカウントした研究データ公開件数を算出する。なお、第一著者が複数の国に所属していたとしても、1 件としてカウントする。

(B) HEPData で公開されているレコードについてであるが、2.3 節で抽出されたレコードのうち半数以上が HEPData で公開されているものであった。HEPData は CERN によって運営されているリポジトリであり、高エネルギー物理学の論文に関連する研究データが公開されている。HEPData で公開されているレコードを含めると分析結果が高エネルギー物理学の分野に偏ってしまうため、これらを除いた調査を行う。DOI の prefix が 10.17182 であるものを HEPData で公開されているレコードとして扱った。

表 1: 研究データ公開件数.

研究データ公開件数カウント方法		研究データ公開件数
HEPData 含	整数カウント	39,128.00
	分数カウント	3,901.81
	第一著者	3,706.00
HEPData 除	整数カウント	5,010.00
	分数カウント	2,566.34
	第一著者	2,968.00
世界		8,175,217.00

本稿では、各要因の選択肢の組み合わせ計 $3 \times 2 = 6$ 通りの方法で研究データ公開件数をカウントする.

3 結果

本章では分析の結果を、全体 (3.1 節), 出版年ごと (3.2 節), 研究データと他リソースの紐付け (3.3 節) に分けて示す.

3.1 全体

表 1 に、研究データ公開件数カウント方法ごとの研究データ公開件数を示す. また、世界の研究データ公開件数 (リソースの種別が Collection, Dataset, Software である DataCite の全レコード件数) を最下行に示す. 論文ではカウント方法によって国や大学ごとの論文件数が大きく異なることが報告されている [5, 6] が、研究データについても同様のことがいえる. 特に、高エネルギー物理学では著者が 1,000 名を超える論文が存在しており、それらの論文の研究データが HEPData で公開されていることから、HEPData を含めた結果では整数カウントと分数カウントの差が大きい. DataCite で公開されている研究データのうち、日本の著者が携わっている研究データ (HEPData を含めた整数カウント) は 0.48% であり、ごく僅かであることがわかる.

3.2 出版年

研究データ公開件数カウント方法ごとの 2010~2020 年の各出版年の研究データ公開件数を図 1 に示す. 世界の各出版年の研究データ公開件数は右軸の目盛りに対応しており、それ以外は左軸の目盛りに対応している. 各レコードの出版年は、レコードのメタデータのフィールド `publicationYear` より

取得した. 2018 年に突起があるものの、HepDATA を含めた整数カウントの場合を除き 2017 年以降は概ね増加傾向にあるといえる.

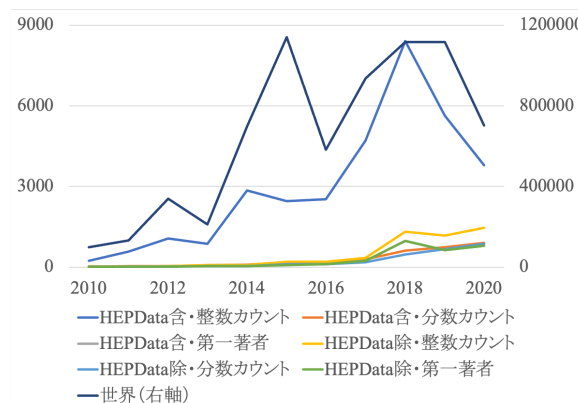


図 1: 出版年ごとの研究データ公開件数.

3.3 研究データと他リソースの紐付け

DataCite の各レコードはフィールド `relatedIdentifiers` によって、関連する他リソースの識別子と紐付けられている. フィールド `relatedIdentifiers` はリストとなっており、各要素には識別子やレコードとの関係の種別 (フィールド `relationType`) が格納されている.

表 2 には、各関係の種別を使用しているレコードの延べ数を示す. 括弧内の数値は表 1 の各研究データ公開件数を使用して、全レコードのうち各関係の種別が使用されている割合を百分率で算出したものである. HEPData を除いた整数カウントで上位 10 件だった関係の種別を対象としている. また、表 3 には、世界の研究データで使用されている関係の種別の上位 10 件、各関係の種別を使用しているレコードの延べ数、各関係の種別が使用されている割合を示している. 1 件のレコードを複数の識別子と紐付けることが可能であることから、100%を超える割合が観察される.

IsSupplementTo は「レコードが関係するリソースを補足している」を表現する際に使用される関係の種別であり、研究データが補足している論文と紐付ける際によく利用されている. 表 2 の HEPData を除いた研究データ公開件数と表 3 を比較すると、日本の大学の著者が公開している研究データは他リソースの補足として使用されている割合が高いことがわかる.

表 2: 各関係の種別を使用している日本の大学の研究データの延べ数。括弧内は各関係の種別が使用している研究データの割合を示す。

関係の種別	HEPData 含				HEPData 除			
	整数カウント	分数カウント	第一著者		整数カウント	分数カウント	第一著者	
IsPartOf	33,292 (85.08)	1,994 (51.12)	1,752 (47.27)		1,965 (39.22)	876.68 (34.16)	1,216 (40.97)	
IsSupplementTo	69,410 (177.39)	3,473 (89.01)	2,328 (62.82)		1,717 (34.27)	859.81 (33.50)	913 (30.76)	
HasPart	1,639 (4.19)	751 (19.24)	1,103 (29.76)		1,639 (32.71)	750.85 (29.26)	1,103 (37.16)	
HasVersion	1,302 (3.33)	531 (13.61)	568 (15.33)		1,302 (25.99)	530.95 (20.69)	568 (19.14)	
IsVersionOf	1,140 (2.91)	528 (13.54)	572 (15.43)		1,140 (22.75)	528.36 (20.59)	572 (19.27)	
IsDocumentedBy	192 (0.49)	97 (2.48)	100 (2.70)		192 (3.83)	96.64 (3.77)	100 (3.37)	
Cites	173 (0.44)	83 (2.13)	57 (1.54)		173 (3.45)	83.16 (3.24)	57 (1.92)	
IsDerivedFrom	121 (0.31)	19 (0.50)	5 (0.13)		121 (2.42)	19.37 (0.75)	5 (0.17)	
References	120 (0.31)	42 (1.07)	7 (0.19)		120 (2.40)	41.91 (1.63)	7 (0.24)	
IsCitedBy	109 (0.28)	73 (1.88)	67 (1.81)		109 (2.18)	73.46 (2.86)	67 (2.26)	

表 3: 各関係の種別を使用している世界の研究データの延べ数。括弧内は各関係の種別が使用している研究データの割合を示す。

関係の種別	研究データの延べ数	
References	103,846,611	(1,270.26)
IsSupplementTo	1,742,136	(21.31)
IsPartOf	1,272,042	(15.56)
HasPart	689,067	(8.43)
HasMetadata	639,934	(7.83)
IsDocumentedBy	486,634	(5.95)
IsPreviousVersionOf	403,529	(4.94)
IsIdenticalTo	321,759	(3.94)
HasVersion	202,507	(2.48)
IsVersionOf	199,062	(2.43)

4 おわりに

本稿では DataCite のレコードのメタデータを使用して、日本の大学における研究データの公開状況を分析した。分析では、カウント方法によって大きく研究データ公開件数が異なること、近年研究データ公開件数は増加傾向にあること、日本の大学の研究データは他リソースの補足として使用されている割合が高いこと等がわかった。

しかし分析には課題がある。2.2 節で述べたとおり、DataCite のレコードの著者数延べ 200,011,615 名のうち、延べ 101,440,687 名に所属機関が記載されていなかった。このことから、実際の日本の大学における研究データ公開件数はさらに多いと思われる。また、本稿では日本の大学に所属する著者を特定するために NISTEP 機関同定プログラムと機関名辞書を使用した。おおそ正しく機関同定を行っていたが、“Tokyo U., ICEPP”を東京大学と認識しないといったような例も観察された。研究データ公開件数をより正確に把握するためには、機関同定プログラムと機関名辞書の評価も必要になってくる。

今後は上記の課題について検討するとともに、紐付けられているリソースを探ることによって研究データを公開する動機を探るなど、詳細な分析を行いたい。

謝辞

本研究は JSPS 科研費若手研究 20K20132 の助成を受けたものである。

参考文献

- [1] 池内有為; 林和弘; 赤池伸一: “研究データ公開と論文のオープンアクセスに関する実態調査”, NISTEP RESEARCH MATERIAL, , No. 268, pp. 1–108, 2017.
- [2] “datacite_dump.20200824”, https://archive.org/details/datacite_dump_20200824, 2020, 最終閲覧日: 2020 年 11 月 23 日.
- [3] “NISTEP 機関同定プログラム公開版の試用実験協力者を募集します”, <https://www.nistep.go.jp/archives/44804>, 2020, 最終閲覧日: 2020 年 11 月 23 日.
- [4] DataCite Metadata Working Group: “DataCite Metadata Schema Documentation for the Publication and Citation of Research Data”, <https://doi.org/10.5438/bmjtbx77>, 2018.
- [5] 中村優文: “共著論文の数え方, 整数カウント・分数カウントとその影響の考察”, 情報の科学と技術, Vol. 70, No. 6, pp. 315–318, 2020.
- [6] 伊神正貫: “論文の生産性分析を考える: 分析者・利用者が確認すべきことと, 分析を実施する上での課題”, STI Horizon, Vol. 4, No. 4, pp. 32–37, 2018.