Vol.2019-CH-119 No.13
2019/2/16

研究データ共有における研究成果および利用者の時系列的分析

中渡瀬 秀一[1]　加藤 文彦[1]　大向 一輝[1]

概要：本研究は研究データ共有が研究活動に与える影響を調査することを目的としている．本稿ではそのような事例として情報学研究データリポジトリ (IDR) に注目し，IDR が配布するデータセット（Yahoo! 知恵袋データ）を用いた研究の成果文献を対象に文献発表量の経年推移や科研費の情報を用いて影響のあった研究分野に関して分析を試みたのでその結果について報告する．

1. はじめに

研究データ共有については，従来よりデータ採取コストダウン・研究成果の検証可能性・研究上の比較基準などの点での動機から研究者個人，研究機関，研究コミュニティなどの各レベルにおいて自発的に進められてきた．日本での現状に目を向けると倉田ら [6] のアンケート調査（2016）に見られるように研究機関が研究者にデータの保管場所を用意し，希望すれば公開も可能となっているとの意見は回答者の 3.3 %に過ぎず，研究データの共有は専ら研究者個人[*1]や分野別リポジトリに委ねられている状況である．

これらとは別の動きとして 2010 年代以降，政策としてのオープンサイエンスにおいてデータ共有が重要視されている．[*2]例えば日本でも内閣府が「国際的動向を踏まえたオープンサイエンスに関する検討会」の報告書 [1] において

> 「オープンサイエンスとは，公的研究資金を用いた研究成果（論文，生成された研究データ等）について，科学界はもとより産業界及び社会一般から広く容易なアクセス・利用を可能にし，（中略）イノベーションの創出につなげることを目指した新しいサイエンスの進め方を意味する。」

との見解を示している．このようにデータ共有に対する期待の高さが政府の認識にも伺われる．

これらの状況を踏まえ，本研究は研究データ共有が研究活動に与える影響を調査・分析することを目的としている．このような影響を調査するためには，共有されたデータに由来する研究成果 (文献) を何らかの手段で観測することが必要である．これには以下の方法がある．

1 データ引用の観測
2 文献本文の調査
3 データ配布主体による記録

それぞれに一長一短があり，1 は文献中にデータ引用が存在すれば引用データベースに記録されるためこの集計は容易で正確である．しかし現状ではデータ引用の習慣は確立されていないため十分に文献を捕捉することは難しい．2 は本文中の記述からデータ利用を判断する方法で，これを自動化する技術はないためコストに難点がある．将来的には言語処理技術による機械化が望まれるが，当面言語処理に伴う精度の低さが解決される見込みは少ない．3 はデータの配布主体が配布先に成果文献の報告を義務付けるものである．この方法が成果文献リストの作成コストとその正確さの点で最も優れておりそのようなデータ配布元を対象にして分析するのが有利である．本稿ではそのようなデータ共有事例として 2 章で説明する情報学研究データリポジトリ (IDR) に注目して分析を行った結果を報告する．以下 2 章では IDR とデータセットについて，3 章では分析方法，4 章では結果と考察，5 章ではまとめを述べる．

2. 情報学研究データリポジトリ（IDR）

IDR は国立情報学研究所 (NII) が提供する分野別リポジトリの一種である．NII は情報学分野における大学共同利用機関[*3]であることから研究資源を研究者に提供することも使命としており，また情報学などの研究分野ではしばしば必要となることが多く作成には多大なコストを要する研

[1] 国立情報学研究所
NII, Chiyoda, Tokyo 101–8430, Japan

[*1] 研究者個人による自発的なデータ類の公開状況の調査としては [7] がある

[*2] 政策としてのデータ共有構想は 1960 年代には既に存在していたが当時の企画（NIST 構想）は実用化には至らなかった．

[*3] 各研究分野における「全大学の共同利用の研究所」として、個別の大学単位では設置や維持が難しい大量の学術データなどを全国の研究者に無償で提供するわが国独自の研究機関．

ⓒ 2019 Information Processing Society of Japan

究資源である大規模データが研究の障害となっていることから，民間企業や研究者から研究用データセットを受け入れて研究者に提供する情報学研究データリポジトリ（IDR）を設置した．

これは冒頭に述べたデータ採取コストの観点によるデータ共有事例に相当する．このリポジトリの特徴には研究者にデータセットの無償提供を行うだけでなく，提供物を使用した研究成果の報告も利用者に求めている点がある．これらは集計されて研究成果リストとなり，これもまたデータセットとして一般に公開されている [4]．すなわち IDR は前述した「データ配布主体による記録」方式を採用しているため分析対象に選んだ．この研究成果リストもまたデータ共有の対象であるから本分析もデータ共有の恩恵を受けて実現した成果である．

この IDR では民間企業からは 2015 年 11 月の時点で 6 企業，13 種類のデータセットを受け入れて提供している [2], [3]．これらの中で最初に配布されたデータセット以下に説明する「Yahoo! 知恵袋データ」である．本分析では時系列を扱うため，最長の提供期間をもつこのデータを対象とした．

2.1 Yahoo! 知恵袋データ

Yahoo! 知恵袋データとは Yahoo!データセット[*4]に含まれるデータセットでヤフー株式会社が運営する Q & A サービス「Yahoo!知恵袋」において解決済みとなった質問と回答等の情報をデータ化したものである．提供時期の異なる 3 種類の版 (表 1) があり現在配布しているのは第 3 版である．本分析では成果リストが存在する第 1 版と第 2 版を対象とする．

表 1 Yahoo! 知恵袋データ

	質問数	回答数	提供開始時期
第 1 版	約 300 万	約 1300 万	2017/04
第 2 版	約 1600 万	約 5000 万	2011/01
第 3 版	約 250 万	約 625 万	2019/01

3. 分析

研究データ共有が研究活動に与える影響を分析するために共有されたデータセットに由来する研究成果 (文献) について次に挙げる 2 つの観点より分析する．

(1)データセットに由来する成果と由来しない関連成果との比較
(2)成果文献が属する研究分野の比較

(1) では「Yahoo!知恵袋」に関する研究文献数に占める「Yahoo!データセット」成果文献の割合を時系列上に比較する．(2) ではデータセットが利用された研究の分野を把握するために「Yahoo!データセット」成果文献の属する研究分野を後述する科研費データベースにより特定し，それらを時系列的かつ分野構造的にマッピングする．この「Yahoo!データセット」成果文献のリストは IDR が Web で公開している「データセット・研究成果一覧 [4]」より収集した．

3.1 分析方法

3.1.1 データセット由来成果と非由来成果との時系列比較

「Yahoo!知恵袋」に関する研究文献はデータセットに由来するものとしないものがあり，前者は成果文献リストから，後者は文献検索サービスの CiNii Articles[*5]を用いて「Yahoo!知恵袋」が含まれる文献を検索して取得した．双方の和集合が文献の全体となる．ただしここでは和文文献だけを対象にしている．これらの文献は次の 3 種のカテゴリに分類される．

- Yahoo!知恵袋データ (第 1 版) の研究成果文献
- Yahoo!知恵袋データ (第 2 版) の研究成果文献
- それ以外の文献

これらの発表時期に応じて 2005 年から 2016 年の時系列上にマッピングする．

3.1.2 成果文献が属する研究分野の比較

3.1.1 で用いた「Yahoo!データセット」成果文献の属する研究分野を特定するため科学研究費助成事業データベース [5] を利用する．このデータには科学研究費助成事業（科研費[*6]）により行われた研究（採択課題）のそれぞれに成果文献が登録されている．また採択課題は属性として研究分野を持つ．つまり双方を照合して，データセット由来の成果文献が含まれる採択課題が存在すればその課題の研究分野がその文献の研究分野となる．このようにして特定された分野を 3.1.1 と同様にして時系列上にマッピングする．科研費の研究分野は階層的 (4 階層) に構造化されており，例えば分野名 (細目) の「メディア情報学」は最上位カテゴリの「総合系」，その下位に続くカテゴリ「情報学」と「計算基盤」の下に位置づけられている．特定された研究分野はこのような分野構造中にもマッピングされる．これにより分野間の近縁度も可視化される．

4. 結果と考察

4.1 データセット由来成果と非由来成果との時系列比較

分析結果を図 1 に示す．「Yahoo!知恵袋」に関する文献は 2005 年から発表が始まり，「Yahoo!データセット」の成果文献 (157 件) は 2008 年から現れた．しかもその数がデータセットに由来しない文献数を超えているためにこの年の全体の発表数は以前の倍以上となっていることが分かる．以後 2011 年までは，データセットに由来する文献が

[*4] NII がヤフー株式会社から提供を受けて研究者に提供しているデータセット
[*5] https://ci.nii.ac.jp/
[*6] https://www.jsps.go.jp/j-grantsinaid/

ⓒ 2019 Information Processing Society of Japan

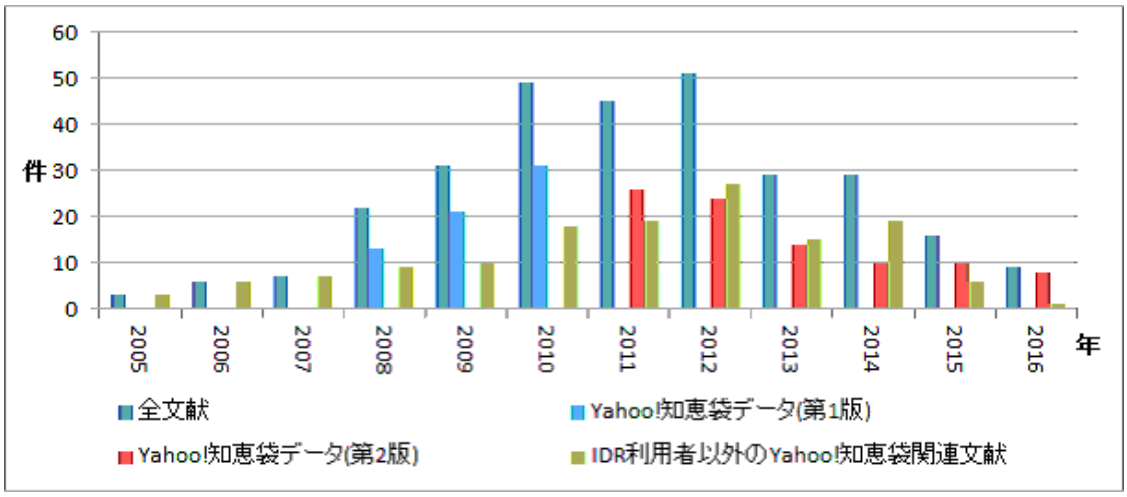

図 1 「Yahoo! 知恵袋データ」を使用した研究成果文献の発表数推移

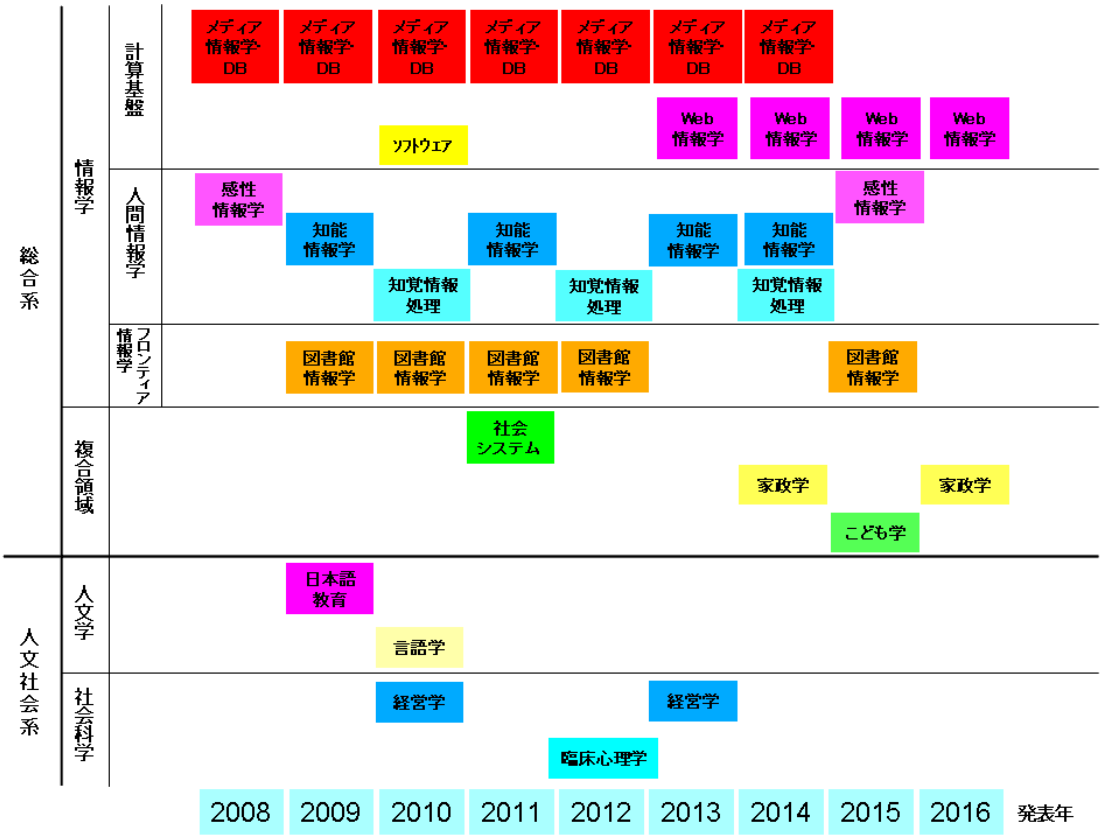

図 2 「Yahoo! 知恵袋データ」を使用した研究成果文献の属する科研費課題の分野分布とその経年推移

それ以外の文献を上回る状態が続いた．これによって全体の文献数も 2008 年以後は急増している．このことはデータ共有による「Yahoo!知恵袋」に関する研究の影響の大きさを示唆している．

Yahoo!知恵袋データ (第 1 版) による成果文献の発表は 2008 年から 2010 年までとなり，以後は第 2 版の成果文献となっている．第 1 版の成果文献数は初年から徐々に増加し最終年にピークを迎えた．ところが第 2 版では初年である 2011 年がピークでありその年も前年を下回った．第 1 版から第 2 版への更新ではデータ量が大幅に拡大されたが，この変化は成果文献数の増加には影響を与えなかった．

4.2 データセット由来成果の分野分布

データセットに由来する成果文献 157 件のうち科研費のデータと照合して研究分野が特定されたものは過半数の 84 件 (54 %) であった．それらの分布を図 2 に示す．IDR が配布するデータセットはその利用目的を情報学に関連する

学術研究に限定している[*7]ので情報学関連の研究分野が多数である．それ以外では情報学に関連する学術研究と判断される分野となるはずであるが，他に見られる複合領域は情報学と同じ上位カテゴリに属しており情報学と近縁の研究分野である．人文学と社会科学は上位カテゴリが同じ人文社会系で相互に近い領域である．つまり「Yahoo!データセット」の場合，その影響は情報学に近い領域や人文社会系に及んでいるといえる．

5. おわりに

本稿では研究データ共有が研究活動に与える影響を把握するために，IDR が配布する「Yahoo!知恵袋データ」に由来する成果文献リストを用いて，成果文献数の経年推移やその研究分野の分布について分析した結果を報告した．その結果，データセットの共有によって研究活動が活性化して成果文献の増大が見られることが判明し，研究分野では情報学とその周辺領域，そして人文社会系の研究分野へも波及していることが示された．現在 Yahoo!データセット以外の IDR が配布するデータセットについても同様の分析を行っており今後，報告する予定である．

「Yahoo!知恵袋データ」の共有では配布先を限定することで成果文献の管理を容易にし，これが分析を可能にしている．オープンサイエンスにおけるデータ共有では研究者のみならず産業界及び社会一般からの広く容易なアクセス・利用が期待されておりこの場合の文献だけに留まらない成果管理が今後の課題となるであろう．

謝辞 本研究にあたり，国立情報学研究所から科研費データの提供を受けた．なお本研究の一部は科研費 挑戦的萌芽研究（課題番号：16K12833）の助成を受けて行われたものである．

[*7] https://www.nii.ac.jp/dsc/idr/yahoo/chiebkr3/Y_chiebukuro.html

参考文献

[1] 内閣府：「国際的動向を踏まえたオープンサイエンスに関する検討会」報告書（2015），入手先 ⟨http://www8.cao.go.jp/cstp/sonota/openscience/150330_openscience_1.pdf⟩（2019.01.01）.
[2] 大山敬三，大須賀智子：情報学研究資源としてのデータセットの共同利用，人工知能学会誌，31(2)，pp. 254–261 (2016)
[3] 大山敬三，大須賀智子：国立情報学研究所における研究用データセットの共同利用，情報管理，59(2)，pp. 105–112 (2016)
[4] データセット・研究成果一覧 (国立情報学研究所データセット共同利用研究開発センター)，入手先 ⟨https://dsc.repo.nii.ac.jp/⟩(2019.01.01).
[5] 科学研究費助成事業データベース，入手先 ⟨https://kaken.nii.ac.jp/⟩(2018.10.01).
[6] 倉田敬子，松林麻実子，武田将季：日本の大学・研究機関における研究データの管理，保管，公開：質問紙調査に基づく現状報告，情報管理，60(2)，pp. 119–127 (2017), https://doi.org/10.1241/johokanri.60.119
[7] 中渡瀬秀一，助成金プロジェクトから見る国内データ成果の現況，情報知識学会誌，27(4)，pp. 370–372 (2017), https://doi.org/10.2964/jsik_2017_044

ⓒ 2019 Information Processing Society of Japan

正誤表

下記の箇所に誤りがございました．お詫びして訂正いたします．

訂正箇所	誤	正
p1,脚注 1	NII	National Institute of Informatics
p1,2 章	また情報学などの研究分野ではしばしば必要となることが多く作成には多大なコストを要する研究資源である大規模データが	また情報学などの研究分野ではしばしば必要となる研究資源である大規模データがその作成コストの高さゆえに
p2, 2 章	これらの中で最初に配布されたデータセット以下に説明する「Yahoo!知恵袋データ」である．	これらの中で最初に配布されたデータセットは以下に説明する「Yahoo!知恵袋データ」である．
p2, 表 1)	第 1 版 約 300 万 約 1300 万 2017/04	第 1 版 約 300 万 約 1300 万 2007/04
p2, 3 章	成果文献の割合	成果文献の数
p3, 4.2 章	4.2 データセット由来成果の分野分布	4.2 成果文献が属する研究分野の比較

ⓒ2019 Information Processing Society of Japan