

データ引用 — 新たな規範への道のり

池内 有為*

キーワード：DOI, FORCE11, データメトリクス, I4OC, Make Data Count (MDC)

1. はじめに

2018年6月号では、研究データを公開する際のライセンスの検討状況について述べた。記事で紹介した質問紙調査の結果、研究者がデータ公開にあたって強く求めている条件は「クレジットの表示」であった。別の調査でも、研究者は公開したデータを「引用せずに利用される可能性」に強い懸念を抱いていることが明らかにされている¹⁾。そこで今回は、公開されている研究データを再利用する際の「データ引用」について取り上げたい。

2. データ引用のいま

論文など出版物の引用は、分野を超えた規範となっている。また、ビッグ3 (APA, シカゴ, MLA) をはじめとする多くのスタイルガイドやデータリポジトリは、出版物と同様に、データやソフトウェアを参考文献リストに記載する場合の書式を示している²⁾。図1に英国のDCC (Digital Curation Centre) によるデータ引用のガイド³⁾から、APAとPANGAEA データリポジトリによる書式の例を示す。

APA
Cool, H. E. M., & Bell, M. (2011). Excavations at St Peter's Church, Barton-upon-Humber [Data set]. doi:10.5284/1000389
PANGAEA
Willmes, S et al. (2009): Onset dates of annual snowmelt on Antarctic sea ice in 2007/2008. doi:10.1594/PANGAEA.701380

図1 データ引用の書式²⁾

しかし、現状では公開されているデータやソフトウェアを再利用する際に、それらに関連する論文を引用することはあっても、データを直接引用している論文は少ない。PMC (旧 PubMed Central) に収録された2017年の論文365,354報のうち、データ引用を含む論文はわずか850報 (全体の0.2%) である²⁾。また、参考文献リストではなく、補足情報³⁾や本文⁴⁾に引用情報やURLを記載する論文が多

い分野もある。

それでは、なぜ関連論文ではなくデータを直接引用する方が良いのか、なぜ参考文献リストに記載する方が良いのか、順に見ていきたい。

3. データ引用の意義

データ引用の背景には、データを論文などの出版物と同様の重要な研究成果とみなすという認識がある。これはOECDやG7をはじめとする各国・地域のオープンサイエンス政策にも共通の認識といえる。象徴的な表現として、英語の文献では“データはファーストクラスのオブジェクト”というフレーズをしばしば目にする。

ところが、データ公開は手間やコストを必要とするのにもかかわらず、論文と同様の業績として認められていない場合が多い。また、データの作成者は、論文の共著者にならない場合もある。そこで論文ではなくデータを引用することによって、(1)データの作成者にクレジットを与えること、(2)データの引用状況を測定して研究の影響度を明らかにして公開のインセンティブとすること — より直截的にいえば、研究助成金の獲得や採用、昇進などにつなげることが期待されている⁵⁾。加えて、データ引用には(3)データへのアクセスを確実にすること、(4)追試や検証を可能にして研究の透明性を向上させること、といったメリットもある。

4. データ引用の標準化と普及の取り組み

こうしたデータ引用のメリットを最大化するために、CODATAとICSTI, RDA (Research Data Alliance) は、データ引用の標準化や普及につとめてきた⁶⁾。これらの機関や出版社といったステークホルダーの協力のもと、国際イニシアティブのFORCE11は、2014年に「データ引用原則の共同声明 (Joint Declaration of Data Citation Principles, JDDCP)」⁷⁾を、2016年には「ソフトウェアの引用指針」⁸⁾を公開した。Elsevierは、1,800タイトルを超える雑誌にJDDCPを採択している。

JDDCPやDCCのガイドなどは、データ引用の測定に対応するために、おおむね次のことを提言している。すなわち、(本文などではなく)参考文献リストにデータ引用を記すこと、著者・タイトル・出版年・出版者/データリポジトリ・ユニークで永続性のあるDOI (Digital Object

*いけうち 有為 筑波大学大学院図書館情報メディア研究科
〒305-8550 茨城県つくば市春日1-2
E-mail: ikeuchi.ui@gmail.com (原稿受領 2018.7.17)

Identifier) を記すことである。

5. 提言による規範化の限界

こうした提言の影響を明らかにするために、Mayo らは Dryad データリポジトリに登録されたデータを用いた論文のデータ引用の状況を調査した⁹⁾。残念ながら、2011年から2014年に出版された1,125報のうち、参考文献リストにDryadのDOIが記載されている論文はわずか6%であった。本文中にDryadのDOIが記載されている論文は75%、どこにも書かれていない論文が20%である。2011年から2014年の経年変化をみると、引用がない論文は31%から15%に減少している。参考文献リストに引用を掲載している論文は5%から8%に増加しているが、このペースでは、2031年まで90%を超えないと指摘している¹⁰⁾。

どうやら、データ引用の書式の整備や提言によってデータ引用を規範とするのには時間がかかりそうである。一方、データ引用の測定や可視化によって、データの探索や選択が容易になれば、また、データ引用が評価されるようになれば、データ引用が広がるのではないだろうか。

6. データ引用の測定

2017年4月、Wikimedia財団やPLOSなど6団体が、引用データの無料公開をめざすI4OC (<https://i4oc.org>)を設立した。I4OCは出版社や学会に協力を呼びかけて、Crossrefに登録されている引用データの公開を進めている。公式サイトによれば、2018年1月現在、Springer NatureやWileyなど154団体が参加しており、雑誌論文の引用データの公開率は活動以前の1%から50%まで増加した。引用データは、CrossrefのREST API (<https://github.com/CrossRef/rest-api-doc>)やOpen Citations Corpus (<http://opencitations.net>)から入手できる。

Crossrefの引用データを用いて、MDC (Make Data Count) プロジェクトはデータのメトリクス (Data Level Metrics) 開発を行っている (<https://makedatacount.org>)。MDCは、2017年5月にCDL (California Digital Library)、DataCite、DataONEが開始した。その前身である Making

Data Count プロジェクトの調査結果を受けて、メトリクスにはデータの引用数、および利用数(閲覧数とダウンロード数)を採用している(図2)。

2018年6月には、CDLとDataONEのリポジトリにメトリクスが実装された。また、MDCはCOUNTERの実務基準を研究データリポジトリに実装するためのガイドをGitHubで公開している¹¹⁾。さらに、より広範な議論を行うべく、RDAにデータ利用メトリクスのワーキンググループ(Data Usage Metrics WG)を立ち上げ、ステークホルダーの参加を呼びかけている。MDCの今後の展開に期待したい。

7. おわりに

引用データがオープン化され、誰もがデータ引用の状況を追跡できる環境が整いつつある(オープンサイエンス!)。データ引用の増加や計量書誌学の進展に期待が高まる反面、メトリクスが実装されることで、インパクトファクターをめぐる不正などの問題が、データについても起きるのではないかと少々心配でもある。ちなみに、論文の被引用数を研究評価に使う場合、批判的な引用や儀礼的な引用も含まれてしまう問題が指摘されている。著者が想像する限り、データを批判的、儀礼的に使うとは考えにくい、はたしてどうだろうか?

註・参考文献

- 注1) 本誌が採用しているSIST 02は、データベース(5.12)やコンピュータプログラム(5.13)を参照する場合の記述方法を定めている。科学技術情報流通技術基準: 参照文献の書き方 [SIST 02]. http://jipsti.jst.go.jp/sist/handbook/sist02_2007/main.htm, (accessed 2018-07-17).
- 注2) 2018年4月、PMCとPubMedは、検索フィルターに「データ引用を含む」などを追加した。検索式は次の通り。(("2017/01/01" [Publication Date]: "2017/12/31" [Publication Date])) AND has data citations [filter]
- 1) 池内有為, 林和弘. 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-オープンサイエンスの課題と展望-. STI Horizon. 2017, vol.3, no.4, p.27-32. <http://doi.org/10.15108/stih.00106>
 - 2) Ball, Alex; Duke, Monica. How to Cite Data Sets and Link to Publications. Digital Curation Centre. 2015, 15p. http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How_to_Cite_Link.pdf, (accessed 2018-07-17).
 - 3) Kafkas, Senay et al. Database citation in supplementary data linked to Europe PubMed Central full text biomedical articles. Journal of Biomedical Semantics. 2015, vol.6, 1. <http://doi.org/10.1186/2041-1480-6-1>
 - 4) Huang, Yi-Hung et al. Citing a data repository: A case study of the Protein Data Bank. PLOS ONE. 2015, vol.10, no.8, e0136631. <http://doi.org/10.1371/journal.pone.0136631>
 - 5) ANDS. Building a Culture of Data Citation. http://www.ands.org.au/_data/assets/pdf_file/0003/383025/data_citation_poster.pdf, (accessed 2018-07-17).
 - 6) CODATA-ICSTI Task Group on Data Citation Standards and Practices; Socha, Yvonne M [ed]. Out of cite, out of mind: the current state of practice, policy, and technology for the citation of data. Data Science Journal. 2013, vol.12, p.CIFCR1-CIFCR75.

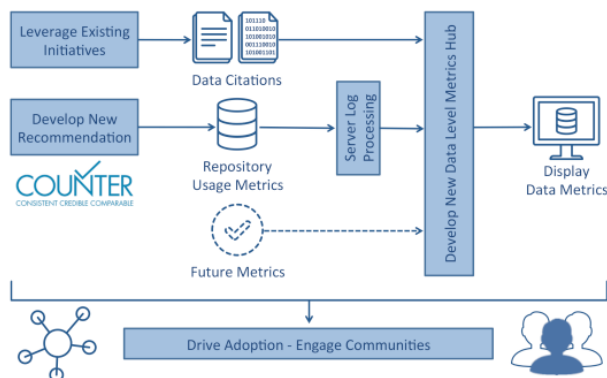


図2 Roadmap&Implementation. Make Data Count. <https://makedatacount.org/roadmap/>

<http://doi.org/10.2481/dsj.OSOM13-043>

- 7) Joint Declaration of Data Citation Principles - FINAL. FORCE11. 2014. <https://doi.org/10.25490/a97f-egykh>
- 8) Smith, Arfon M. et al. Software citation principles. PeerJ Computer Science. 2016, vol.2, e86. <https://doi.org/10.7717/peerj-cs.86>
- 9) Mayo, Christine et al. The location of the citation: Changing practices in how publications cite original data in the Dryad Digital Repository. International Journal of Digital Curation. 2016, vol 11, no.1, p.150-155. <http://doi.org/10.2218/ijdc.v11i1.400>
- 10) Hull, Elizabeth. "The Location of the Citation: Are Data Citation Recommendations Having an Effect?". DataCite Blog. 2016-3-28. <https://doi.org/10.5438/f17b-45vz>
- 11) Make Data Count. Implementing the COUNTER Code of Practice for Research Data in Repositories. GitHub. 2018. <https://github.com/CDLUC3/Make-Data-Count/blob/master/getting-started.md>, (accessed 2018-07-17).

Series: Current trend of open science. Data citation—Road to a new norms. Ui IKEUCHI (Graduate School of Library, Information and Media Studies, University of Tsukuba, 1-2 Kasuga, Tsukuba-shi, Ibaraki 305-8550)

Keywords: DOI / FORCE11 / Data Metrics / I4OC / Make Data Count (MDC)