【予稿集】

英語版 Wikipedia における DOI リンクの初出時点の分析
― 研究分野を中心に ―

吉川 次郎 *, 高久 雅生 **
* 筑波大学大学院図書館情報メディア研究科
** 筑波大学図書館情報メディア系
*jiro@slis.tsukuba.ac.jp, **masao@slis.tsukuba.ac.jp

英語版 Wikipedia における DOI リンクを通じた学術文献の参照記述を対象に、初出時点および研究分野に着目した分析を行なった。その結果、(1) 参照記述は 2002 年に初めて追加され、2007 年から急増している、(2) 参照記述の追加は近年になればなるほど盛んになっている、(3) Scopus 収録雑誌に対する論文参照率は全体で 1.2%で、研究分野では「Multidisciplinary」が 3.9%で最も高かった。

Analyses of First Appearance of DOI Links on English Wikipedia
– From the View Point of Research Fields –

Jiro KIKKAWA*, Masao TAKAKU**
*Graduate School of Library, Information and Media Studies, University of Tsukuba
**Faculty of Library, Information and Media Science, University of Tsukuba

1. はじめに

今日、学術情報流通の電子化を背景に、ウェブ上で学術情報の参照が行われている。これらの参照は、従来の学術論文における引用・被引用を代替・補完しうるもの、すなわち、代替的な評価指標 (Altmetrics) として注目を集めている。そのデータソースの 1 つとして、誰でも編集可能なオンラインのフリー百科事典である「Wikipedia」がある。

世界最大の DOI (Digital Object Identifier、デジタルオブジェクト識別子) [1] の登録機関である Crossref は、2015 年時点でのアクセスログの分析を通じて、Web of Science や Scopus 等の学術文献データベースに次いで Wikipedia が 5 番目にアクセスの多い参照元であることを指摘している [2]。このことから、Wikipedia において、DOI リンクを通じた学術文献の参照記述が存在し、閲覧者は、これらのハイパーリンクを通じて学術文献にアクセスしている。

筆者らは、これまでに日本語、中国語、英語版 Wikipedia における DOI リンクの分析を行なってきた。その結果として、日本語版および中国語版における DOI リンクの参照記述は、その大部分が、英語版の翻訳 (コピー) を通じて記述されたものであることが明らかになった [3]。この結果を踏まえ、本研究では、英語版 Wikipedia を分析対象とする。

Wikipedia 上での学術文献の参照に関する研究は、複数行なわれているものの、それぞれの学術文献の参照記述がいつから追加されているのかを明らかにする研究や、研究分野ごとの参照状況に着目した研究は、筆者らの研究 [4,5] を含め、僅かである。

初出時点に関しては、Halfaker ら [6] による、英語版およびオランダ語版 Wikipedia 上での DOI、PubMed、ISBN、arXiv の識別子の参照記述を対象とした分析事例が存在する。また、特定の研究分野における学術文献に関する分析事例としては、Thelwall [7] が英語版 Wikipedia における天文学分野を対象とした分析を行なっている。これらの研究とは異なり、本研究は、DOI リンクを通じた学術文献の参照記述の精緻な分析を目指し、複数の

研究分野を対象とした分析を行なう。

本研究では、英語版 Wikipedia 上の DOI リンクを対象に、初出時点の分析による参照記述の追加件数の経年変化と、研究分野ごとの参照状況を明らかにする。具体的には、以下の 3 つのリサーチクエスチョン (RQ) に答えることを目指す:

RQ1 参照記述は何年頃から追加されているのか?

RQ2 参照記述は近年になるにつれて盛んに追加されているのか?

RQ3 どの研究分野の雑誌や論文が、どれくらい参照されているのか?

2. 対象と方法

2.1 分析対象

分析対象は、2017 年 3 月 5 日までに英語版 Wikipedia に追加された DOI リンクのうち、Scopus 収録雑誌の論文に対する参照記述である。具体的には、Crossref DOI かつ「Scopus タイトルリスト[1]」収録雑誌の論文が分析対象である。研究分野単位での分析のために、同リストの「All Science Journal Classification (ASJC) Code」における最も大きな分野区分である「Health Science」、「Life Science」、「Multidisciplinary」、「Physical Science」、「Social Science」の 5 分野の分類を用いる。

2.2 データセットの構築

まず、英語版 Wikipedia の DOI リンクの参照記述を含むページを特定するために、2017 年 3 月 5 日時点の英語版 Wikipedia のダンプデータ[2]を用いて、百科事典記事を意味する「標準名前空間」における DOI リンクの参照記述および当該記述を含むページを取得した。

次に、Wikipedia の提供する Web API である「API:Revisions[3]」を用いて DOI リンクの参照記述を含むページの編集履歴の一覧を取得し、編集日時が最古のものから順に、各編集時点において参照されているハイパーリンクを「API:Parsing wikitext[4]」により取得した。これらのデータにより、あるページ上で任意の DOI リンクの参照記述に対する初出時点を特定できる。このとき、DOI リンクから DOI 名を取得したうえで、アンエスケープを施した文字列を大文字・小文字の区別をせずに比較した。したがって、同一ページ上の複数箇所に同一の DOI リンクが記述されている場合は、最も古い参照記述のみが分析対象となる。

最後に、Crossref REST API[5]を用いて、Crossref DOI のメタデータを取得し、Scopus タイトルリストに収録されている雑誌か否か、どの研究分野の雑誌であるかの判定を行なう。同 API が返戻するメタデータには、たとえば論文の場合、タイトルや雑誌名等の情報に加え、ISSN が含まれるため、Scopus タイトルリストとの照合には ISSN を用いた。

以上の手順を通じて構築したデータセットの概要を表 1 に示す。表 1 から、英語版 Wikipedia 上での DOI リンクを通じた参照記述の初出時点は 1,121,604 件であり、これらのうち、Crossref DOI は 1,099,788 件で、初出時点全体の 98.1%に相当する。「Crossref DOI かつ Scopus 収録あり」、すなわち、Scopus 収録雑誌の論文に対する DOI リンクを通じた参照記述は 997,608 件であり、初出時点全体の 88.9% (異なり DOI 名では 88.3%) に相当する。以上から、英語版 Wikipedia における DOI リンクを通じた参照記述の大部分は、Crossref DOI かつ Scopus 収録雑誌の論文である。

2.3 分析方法

「Crossref DOI かつ Scopus 収録あり」の条件を満たす延べ 997,608 件の参照記述の分析を行なった。

1 点目は、初出時点の件数に関する経年変化の分析である。初出時点を 1 年単位で集計し、ある年に何件の参照記述が追加されたかを分析するとともに、累積件数を算出した。

2 点目は、Scopus の研究分野の分類を用いた集計である。ここでは研究分野ごとに、雑誌および論文の参照率を算出した。Scopus 収録雑誌およびその掲載論文のなかでも、Crossref DOI を通じて

[1] https://www.elsevier.com/?a=91122
[2] https://dumps.wikimedia.org/backup-index.html
[3] https://www.mediawiki.org/wiki/API:Revisions
[4] https://www.mediawiki.org/wiki/API:Parsing_wikitext
[5] https://api.crossref.org/

表 1: 分析対象の概要

項目/条件	初出時点全体	Crossref DOI	Crossref DOI かつ Scopus 収録あり	Crossref DOI かつ Scopus 収録なし
延べ DOI 名	1,121,604	1,099,788	**997,608**	102,180
割合 (%)	–	100.0	**90.7**	9.3
異なり DOI 名	759,112	742,140	**670,233**	71,907
割合 (%)	–	100.0	**90.3**	9.7
異なりページ	221,157	219,500	198,067	61,158

参照可能なものに限り、割合を算出した。

雑誌の参照率は、英語版 Wikipedia で参照されている論文の掲載誌を Print ISSN を用いて特定し、Scopus 収録雑誌における割合を算出した。

論文の参照率は、まず、Print ISSN を用いて、すべての Crossref DOI のうち、Scopus 収録雑誌の論文数を算出した。さらに、これらの論文のうち、英語版 Wikipedia 上で参照されている論文の割合を算出し、論文の参照率とした。

Scopus の分類では、ある 1 つの雑誌が複数の研究分野に分類される場合があるが、82.0% (30,219 誌) は 1 分野、16.7% (6,158 誌) は 2 分野に分類されており、大部分が 2 分野以内に分類されることを踏まえ、複数分野に分類されている場合も区別せずに集計・分析を行なった。

3. 分析結果

3.1 初出件数の経年推移

表 2 より、DOI リンクを通じた学術文献の参照記述が初めて追加されたのは 2002 年である。

経年変化としては、初期の数年間での追加件数は僅かであるが、2007 年から件数が急増している。その後、2010 年から 2013 年には横ばい または 減少傾向にあるが、2014 年には再び増加傾向が見られ、以降は年間 10 万件以上の追加がコンスタントに行なわれている。

表 2: 参照記述の追加件数の経年変化

年/項目	初出件数	割合 (%)	累積件数
2002 年	1	0.0	1
2003 年	2	0.0	3
2004 年	64	0.0	67
2005 年	313	0.0	380
2006 年	3,779	0.4	4,159
2007 年	55,544	5.6	59,703
2008 年	99,317	10.0	159,020
2009 年	99,526	10.0	258,546
2010 年	87,240	8.7	345,786
2011 年	71,526	7.2	417,312
2012 年	71,605	7.2	488,917
2013 年	72,074	7.2	560,991
2014 年	124,313	12.5	685,304
2015 年	156,775	15.7	842,079
2016 年	133,853	13.4	975,932
2017 年	21,676	2.2	997,608
全体	**997,608**	**100.0**	**997,608**

3.2 研究分野ごとの雑誌の参照率、論文の参照率

Scopus の全収録雑誌 36,832 件のうち、Crossref DOI を通じて参照可能なものに限定した場合の雑誌 21,612 件と、その掲載論文 57,429,821 件における雑誌の参照率と論文の参照率を表 3 に示す。

表 3 から、雑誌の参照率は「Life Science」(96.4%) が最も高く、「Multidisciplinary」(69.8%) が最も低い。論文の参照率は「Multidisciplinary」(3.9%) が最も高く、「Physical Science」(0.8%) が最も低い。全体で見ると、雑誌の参照率は 81.8%、論文の参照率は 1.2%である。

4. 考察と今後の課題

本研究では、英語版 Wikipedia 上の DOI リンクを対象に、初出時点の分析による参照記述の追加件数の経年変化と、研究分野ごとの参照状況を明らかにするために、1 章で示した RQ に答えるための分析を行なった。

RQ1 は、表 2 の結果から、2002 年に参照記述の追加が初めて行なわれ、2007 年から急増している。

表 3: 研究分野ごとの雑誌の参照率、論文の参照率

分野名/項目	Scopus			DOI リンクを通じた参照記述			
	全収録	DOI 参照可能な		異なり	参照率	異なり	参照率
	雑誌数	雑誌数	論文数	雑誌数	(%)	論文数	(%)
Health Science	13,819	7,152	21,747,085	5,792	81.0	222,774	1.0
Life Science	6,809	4,678	14,771,374	4,510	96.4	300,731	2.0
Multidisciplinary	115	63	1,089,733	44	69.8	42,484	3.9
Physical Science	12,263	7,277	22,432,450	6,197	85.2	176,813	0.8
Social Science	10,905	7,417	10,661,135	5,605	75.6	105,869	1.0
全体	36,832	21,612	57,429,821	17,683	81.8	670,233	1.2

RQ2 は、表 2 の結果から、近年になるにつれて盛んに追加されていると言える。

表3から、雑誌の参照率は「Life Science」(96.4%) が最も高く、「Multidisciplinary」(69.8%) が最も低い。論文の参照率は「Multidisciplinary」(3.9%) が最も高く、「Physical Science」(0.8%) が最も低い。全体で見ると、雑誌の参照率は 81.8%、論文の参照率は 1.2%である。

今後の課題は、(1) 参照記述を追加する編集者の分析、(2) 学術文献が公表・出版されてから Wikipedia 上で参照されるまでのタイムラグの分析、(3) 論文間での引用・被引用と Wikipedia における参照との差異の分析を行なうことである。

参考文献

[1] Japan Link Center. “DOI ハンドブック”. ジャパンリンクセンター (JaLC). https://doi.org/10.11502/DOI_Handbook, (参照 2018-05-30).

[2] Bilder, Geoffrey. “Geoffrey Bilder: Strategic Initiatives Update”. SlideShare. 2015-11-23. http://www.slideshare.net/CrossRef/geoffrey-bilder-crossref15, (参照 2018-05-30).

[3] Jiro, Kikkawa; Masao, Takaku; Fuyuki, Yoshikane. “DOI Links on Wikipedia: Analyses of English, Japanese, and Chinese Wikipedias”. Lecture Notes in Computer Science. 2016, vol.10075, p.369-380. https://doi.org/10.1007/978-3-319-49304-6_40, (参照 2018-05-30).

[4] 吉川次郎, 高久雅生. “Wikipedia における DOI リンクの経年変化の予備的分析”. 第 22 回情報知識学フォーラム. 東京, 2017-12-02. 情報知識学会誌. 2017, vol.27, no.4. p.329-336. https://doi.org/10.2964/jsik_2017_036, (参照 2018-05-30).

[5] 吉川次郎, 高久雅生. “Wikipedia 上の J-STAGE コンテンツの分析: 研究分野を中心に”. 情報メディア学会第 15 回研究大会. 茨城, 2016-06-25. 第 15 回情報メディア学会研究大会発表資料. 2016, p.39-42. http://hdl.handle.net/2241/00143015, (参照 2018-05-30).

[6] Halfaker, Aaron; Taraborelli, Dario. “Research:Scholarly article citations in Wikipedia”. Meta, a Wikimedia project coordination wiki. https://meta.wikimedia.org/wiki/Research:Scholarly_article_citations_in_Wikipedia, (参照 2018-05-30).

[7] Thelwall, Mike. “Does Astronomy research become too dated for the public? Wikipedia citations to Astronomy and Astrophysics journal articles 1996-2014”. El Profesional de La Informacion. 2016, vol.25, no.6, p.893-900. https://doi.org/10.3145/epi.2016.nov.06, (参照 2018-05-30).