Tulips R
University of Tsukuba Repository

日本語版WikipediaにおけるDOIリンクの予備的分析

著者	吉川 次郎, 高久 雅生, 逸村 裕
著者別名	KIKKAWA Jiro, TAKAKU Masao, ITSUMURA Hiroshi
雑誌名	情報知識学会誌
巻	25
号	2
ページ	160-165
発行年	2015
権利	情報知識学会
その他のタイトル	Preliminary Analyses of DOI Links on Japanese Wikipedia
URL	http://hdl.handle.net/2241/00125064

第 23 回年次大会予稿

日本語版 Wikipedia における DOI リンクの予備的分析
Preliminary Analyses of DOI Links on Japanese Wikipedia

吉川 次郎[1]* 高久 雅生[2] 逸村 裕[3]
Jiro KIKKAWA[1]*, Masao TAKAKU[2], Hiroshi ITSUMURA[3]

[1] 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: jiro@slis.tsukuba.ac.jp
[2] 筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: masao@slis.tsukuba.ac.jp
[3] 筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba
〒 305-8550 茨城県つくば市春日 1-2
E-mail: hits@slis.tsukuba.ac.jp
*連絡先著者 Corresponding Author

本研究では，日本語版 Wikipedia に含まれる DOI リンクの予備的分析を行った．2015 年 3 月時点での日本語版 Wikipedia には 28,546 件の DOI リンクがあり，そのうち標準名前空間ページに含まれる 27,201 件について，DOI 登録機関は CrossRef が 97%，JaLC が 2%であった．日本国外の大手出版社が多く，雑誌タイトルレベルでは，Nature，Science，PNAS などの自然科学分野の有力誌が多く含まれていた．また，日本国内の出版者の DOI リンクが含まれていた．

The authors extracted and analyzed DOI links from external links in Japanese Wikipedia. As of March 2015, there were 28,546 DOI links in all pages and 27,201 in main namespace pages. In terms of Registration Agencies, 97% of DOI Links were registered by CrossRef, and 2% by JaLC. From the aspect of Journal Titles, most journals were from scientific field. Some were Japanese journals.

キーワード: Wikipedia，Digital Object Identifier(DOI)，学術情報流通
Keyword: Wikipedia, Digital Object Identifier(DOI)，Scholarly Communication

1 はじめに

学術情報流通において，学術論文をはじめとする研究成果は不可欠な存在であり，これを誰もが障壁なく利用できるようにすることの重要性がオープンアクセスの文脈において盛んに論じられてきた．さらに近年では，最終的な成果物としての学術論文のみでなく，研究の過程で生じるデータに関してもオープン化の機運が高まっている．

誰もがウェブ上で学術情報を自由に利用できる環境が提供されることは，従来の学術情報の利用者として認識されてきた研究者や専門家だけではなく，一般市民を含む，より広範な人々が学術情報を利用しうることを意味する．さらに近年,「オープンサイエンス」として，研究成果や研究データの利活用を促進する動きがあることから，ウェブで利用可能な学術情報の量およびそれを利活用する動きは今後も拡大を続けることが考えられる．このように，オープンなウェブと学術情報が結びつくことで，従来と異なる学術情報の利用が生じると考えられる．

2014 年時点で日本におけるパソコンからの利用者数の第 8 位 [1] である Wikipedia は，誰でも編集

できるフリー百科事典であり，学術的な内容を含む多様な項目が存在する．Wikipedia の本文中の記述に関する出典として学術情報への外部リンクが含まれている．また，DOI(Digital Object Identifier) の世界最大規模の登録機関である CrossRef の報告では，CrossRef が登録した DOI に関して，参照元のうち 8 番目に大きなウェブサイトが Wikipedia であり，実際に Wikipedia の利用者が DOI リンクをクリックして閲覧していることを指摘している [2]．このことから，Wikipedia がウェブと学術の世界を橋渡しする役割を果たしていることが分かる．

Wikipedia の外部リンクを分析した事例としては，Nielsen が英語版 Wikipedia の外部リンクに含まれる学術情報に着目し，当該論文の Journal Citation Reports でのインパクトファクター値との関係性を調査した研究 [3] がある．日本語版 Wikipedia については，佐藤らによるリンク切れの状況に着目した分析 [4] がある．

しかし，日本語版 Wikipedia の外部リンクに含まれる学術情報に着目した調査は管見の限り見当たらない．また，DOI についても，システム自体の仕組みや付与対象に関する解説記事は見られるものの，DOI がどのような場所で，どのような学術情報を繋いでいるのかについての分析事例は見当たらない．そこで本研究は日本語版 Wikipedia のダンプデータをもとに，外部リンク内の DOI リンクを分析した．

2 「DOI」と「DOI リンク」

DOI とは，コンテンツの電子データに付与される国際的な識別子であり，解決可能，持続的，相互運用可能なリンクを提供するための仕組みである．

DOI は，図 1 に示すように「10.」で始まる Prefix，「/」(スラッシュ)，Suffix，から構成される．

DOI を「`http://dx.doi.org/`」(または「`http://doi.org/`」) の後方に加えることで URL として機能し，当該コンテンツの URL にリダイレクトされる．本研究では，この URL を通じたハイパーリンクを「DOI リンク」と定義する．

DOI の登録は RA(Registration Agencies，DOI 登録機関) を通じて行われる．世界最大規模の RA は CrossRef であり，日本国内においては Japan Link Center(以下，JaLC とする) が唯一の RA である．2015 年 4 月時点で，CrossRef によって登録された DOI は約 7,300 万件，JaLC によって登録された DOI は約 300 万件である [5] [6]．他の主な RA としては，Data Cite，mEDRA，OPOCE がある．

DOI の例
10.2964/jsik.21_06
DOI リンクの例
http://dx.doi.org/10.2964/jsik.21_06
（Prefix: 10.2964 Suffix: jsik.21_06 ）

図 1: 「DOI」と「DOI リンク」の例

3 対象・方法

3.1 分析対象

本研究では，2015 年 3 月 13 日時点の日本語版 Wikipedia のダンプデータのうち，外部リンクが記述された「externallinks.sql」と項目名ごとのデータが記述された「pages.sql」を分析に使用した．

分析対象となる DOI リンクの抽出については，内部結合により「pages.sql」の項目名，名前空間 (page_namespace) とともに，外部リンク群から「dx.doi.org」または「doi.org」を含むリンクを抽出した．抽出した 28,552 件から非 DOI リンクである 6 件を除去した DOI リンク 28,546 件のうち，百科事典の記事を意味する，名前空間が「0」(以下，標準名前空間とする) である 27,201 件を主な分析対象とする．

3.2 分析方法

標準名前空間における 27,201 件の DOI リンクについて，doiRA [7] と呼ばれる API を用いて RA のデータを取得する．さらに，CrossRef REST API(以下，REST API とする) [8] を用いてデータを取得する．以下，それぞれの API を実行して得られるデータの具体例を示す．

{“DOI”: “10.2964/jsik.21_06”,
“RA”: “CrossRef”}

図 2: doiRA による RA の取得例

doiRA は，任意の DOI について，RA のデータを取得する API である．図 2 は「DOI:10.2964/jsik.21

_06」の RA を取得した例で，RA は「CrossRef」である．doiRA は CrossRef 以外が登録した DOI についても RA のデータを取得できる．

{“member”:“http:\/\/id.crossref.org\/member\/1527”, “name”:“Japan Society of Information and Knowledge”, “prefix”: “http:\/\/id.crossref.org\/prefix\/10.2964”}

図 3: REST API の prefixes を取得した例

REST API の prefixes により，Prefix のデータを取得できる．図 3 は「Prefix:10.2964」の prefixes を取得した例である．このとき，name の値から「Japan Society of Information and Knowledge」が登録者 (以下，Registrant とする) であることが分かる．ただし，REST API の prefixes は RA が CrossRef である DOI のみに対応しており，それ以外の場合は Registrant を取得することができない．

{...“container-title”:[“Journal of Japan Society of Information and Knowledge”, “Joho Chishiki Gakkaishi”]...}

図 4: REST API の works を取得した例

Journal Title については，REST API の works によって取得が可能である．図 4 は「Prefix:10.2964, Suffix:jsik.21_06」の情報を取得した例である．このとき，container–title の値より，ジャーナル名 (以下，Journal Title とする) は「Journal of Japan Society of Information and Knowledge, Joho Chishiki Gakkaishi」である．REST API の works は RA が CrossRef である DOI のみに対応しており，それ以外の場合は Journal Title を取得することができない．

4 分析結果

4.1 名前空間ごとの集計

分析対象の DOI リンク 28,546 件を名前空間ごとに集計した結果を表 1 に示す.「のべ DOI リンク数」は重複を含む DOI リンクの数,「異なりページ」は重複を除いたページ数,「異なり DOI リンク数」は重複を除いた DOI リンク数を指す.

表 1: 名前空間ごとの集計 (n=28,546)

名前空間	のべ DOI リンク数	異なりページ数	異なり DOI リンク数
0	27,201	9,135	24,071
1	69	48	66
2	568	122	558
3	45	7	44
4	7	5	6
5	2	2	2
6	2	1	2
10	634	632	623
11	2	2	2
102	14	4	14
103	2	1	2
全体	28,546	9,959	24,599

DOI リンクが含まれている名前空間は,「0 標準」,「1 ノート」,「2 利用者」,「3 利用者–会話」,「4 Wikipedia」,「5 Wikipedia–ノート」,「6 ファイル」,「10 Template」,「11 Template–ノート」,「102 プロジェクト」,「103 プロジェクト–ノート」である．ノートページや利用者ページなど，標準名前空間ページ以外においても DOI リンクが含まれている．DOI リンクが最も多く含まれているのは標準名前空間ページであり，27,201 件の DOI リンクを含んでいる．次いで，Template ページ，利用者ページ，ノートページの順に多い．

脚注のファーイーストリサーチ社の引用はやめたほうがよいと思います。「特命リサーチ200X」という番組内の架空の調査会社だといわないと、そのページだけ見ると、実在する社のような印象を受けてしまいます。また、そのページの報告者も俳優が演じている架空の人物なので、その点もまぎらわしいです。もっと適切な引用先がみつかればよいのですが……。Google Scholarで調べてもあまり引っかかりませんでした。

Seagal, N. L. (2000). New breast cancer research: Mothers and twins. *Twin Research and Human Genetics*, *3*, 118-122. DOI: 10.1375/twin.3.2.118

Seagal, N. L. (2001). Twin assortment. *Twin Research and Human Genetics*, *4*, 122-123. DOI: 10.1375/twin.4.2.122

Shaz 2007年8月15日 (水) 03:32 (UTC)

図 5: ノートページでの DOI リンクを用いた議論の例 (下線は筆者による)．出典は参考文献 [9]

また，図 5 のように，本文中での記述の出典に関してノートページで議論が行われる際に，DOI リンクが用いられる事例がある．

4.2 RA ごとの集計

標準名前空間ページに含まれる DOI リンクについて，RA ごとの集計結果を表 2 に示す．

最も件数の多い RA は CrossRef であり，26,273 件 (約 97%) と全体の大部分を占めている．次いで多い RA は JaLC であり，518 件 (約 2%) である．CrossRef は学術論文などの学術的なコンテンツに DOI を付与する RA であるため，日本語版 Wikipedia における DOI リンクの大部分は学術論文などの引用である．JaLC は日本国内で発行された学術的なコンテンツに DOI を付与する RA であることから，日本国内で刊行された学術論文の引用も行われている．

上記以外の RA として，Data Cite，mEDRA，OPOCE，Public がある．Data Cite は研究データセットに DOI を付与する RA であるため，少数ではあるものの，日本語版 Wikipedia において研究データの引用が行われているといえる．

その他の結果としては，「DOI does not exist」，「Invalid DOI」，「Error」があり，それぞれ，存在しない DOI，無効な DOI，それ以外のエラー，の場合に得られる結果である．「DOI does not exist」は Suffix の値の記述に誤りがある場合や，RA 側の登録内容に問題があることが原因と考えられる．「Invalid DOI」は Prefix に「doi:」のような不要な文字列が含まれている場合である．「Error」は 404 のステータス・コードが返される場合である．これらの結果から，日本語版 Wikipedia には存在しない DOI リンクや無効な DOI リンクが含まれている．

表 2: RA ごとの集計 (n=27,201)

RA	Count	備考
CrossRef	26,373	
Japan Link Center	518	
Data Cite	11	
mEDRA	5	
OPOCE	2	
Public	6	
DOI does not exist	186	存在しない DOI
Invalid DOI	99	無効な DOI
Error	1	上記以外のエラー

4.3 Prefix ごとの集計

標準名前空間ページに含まれる DOI リンクについて，Prefix ごとの集計結果を表 3 に示す．なお，「Wiley–Blackwell」のように Registrant と Prefix は 1 対多の関係になる場合がある．

上位の Registrant は日本国外の大手出版社である．1 位「Elsevier」，2 位「Springer」，4 位「Nature」，5&6 位「Wiley–Blackwell」，13 位「Informa UK Limited」は商業出版社系，3 位「ACS」，8 位「AAAS」，10 位「PNAS」，11 位「APS」，12 位「Oxford University Press」，14 位「ASBMB」，15 位「EDP Sciences」は学協会系の有力誌である．

表 3: Prefix ごとの集計 (上位 15 件，n=27,201)

No	Prefix	Registrant	Count
1	10.1016	Elsevier BV	4,398
2	10.1007	Springer Science +Business Media	1,707
3	10.1021	ACS	1,651
4	10.1038	Nature Publishing Group	1,424
5	10.1002	Wiley–Blackwell	1,292
6	10.1111	Wiley–Blackwell	1,235
7	10.1086	University of Chicago Press	1,112
8	10.1126	AAAS	853
9	10.2307	JSTOR	649
10	10.1073	PNAS	573
11	10.1103	APS	550
12	10.1093	Oxford University Press	527
13	10.108	Informa UK Limited	421
14	10.1074	ASBMB	335
15	10.1051	EDP Sciences	317

その他，表 3 には登場しないものの，特筆すべき点として，オープンアクセスメガジャーナルの「PLoS」が 20 位 (241 件，約 0.9%) である．日本国内の出版者は，「Tokyo Geographical Society(東京地学協会)」が 35 位 (132 件，約 0.5%)，「Japanese Society of Fisheries Science(日本水産学会)」が 37 位 (120 件，約 0.4%)，「Japan Society for Bioscience, Biotechnology, and Agrochemistry(日本農芸化学会)」が 38 位 (112 件，0.4%) である．

表 4: Journal Title ごとの集計 (上位 15 件，n=27,201)

No	Journal Title	Category name	Count
1	Nature	Multidisciplinary	902
2	Science	Multidisciplinary	839
3	PNAS	Multidisciplinary	569
4	JACS	CHEMISTRY	564
5	The Astrophysical Journal	SPACE SCIENCE	463
6	Journal of Biological Chemistry	BIOLOGY & BIOCHEMISTRY	327
7	Astronomy and Astrophysics	SPACE SCIENCE	278
8	The Astrophysical Journal	SPACE SCIENCE	251
9	Biochemistry	BIOCHEMISTRY	231
10	European Journal of Biochemistry	BIOCHEMISTRY	214
11	Physical Review Letters	PHYSICS	206
12	Archives of Biochemistry and Biophysics	BIOCHEMISTRY	190
13	Icarus	SPACE SCIENCE	176
14	New England Journal of Medicine	CLINICAL MEDICINE	165
15	The Journal of Organic Chemistry	CHEMISTRY	155

4.4 Journal Title ごとの集計

標準名前空間ページに含まれる DOI リンクについて，Journal Title ごとの集計結果を表 4 に示す．

上位 15 件について THOMSON REUTERS の Master Journal List [10] にある「SOURCE PUBLICATION DOCUMENTS」との照合を行ったところ，すべて「Science Citation Index Expanded Source Publication」に収録されていることが分かった．このことから，上位 15 件は Journal Citation Reports での Science 分野のジャーナルであることが分かる．さらに詳細情報を得るため，THOMSON REUTERS が提供する SCIENCEWATCH の「JOURNAL LIST FOR ESSENTIAL SCIENCE INDICATORS」[11] との照合結果を Category name として示す．

Journal Title ごとに見ると,「Nature」(902 件, 約 3%) が最も多く,「Science」(839 件, 約 3%),「PNAS」(569 件, 約 2%) が続く．これらはいずれも Category name が「Multidisciplinary」である．

Category name について件数の多い順に見ると,「Multidisciplinary」(2,310 件),「SPACE SCIENCE」(1,168 件),「CHEMISTRY」(719 件),「BIOCHEMISTRY」(635 件),「BIOLOGY & BIOCHEMISTRY」(327 件),「PHYSICS」(206 件),「CLINICAL MEDICINE」(165 件) である．この結果から，学際分野のほか，宇宙科学，化学，生化学，生物学，物理，臨床医学分野が多く含まれているといえる．ただし，Journal Title は CrossRef 側に登録されたデータをそのまま出力した結果であるため，1 つの Journal Title が複数に分かれてしまっている可能性がある．

なお，表 4 から除外した項目として，REST API での works の取得結果が「Resource not found.」のものが 861 件ある．表 2 での CrossRef 以外の RA による DOI が含まれ，JaLC の DOI 518 件すべてが該当する．

5 考察と今後の課題

本研究では，日本語版 Wikipedia の外部リンクに含まれる DOI リンクの分析を行った．2015 年 3 月時点で，日本語版 Wikipedia 全体に DOI リンクは約 28,546 件あり，標準名前空間，Template，利用者ページに多く含まれていた．

標準名前空間ページの 27,201 件の DOI リンクのうち，97%が CrossRef，2%が JaLC によって登録されたものであった．CrossRef の DOI リンクについて，大部分は日本国外の大手商業出版社や学協会のコンテンツである．ジャーナルとしては自然科学分野が多く，Nature や Science などの総合誌のほか，宇宙科学，化学，生化学，生物学，物理，臨床医学分野であった．

これらの結果は，日本語版 Wikipedia において，従来の研究者や専門家による学術情報流通とあまり変わりのない学術情報が利用されていることを示すものである．その一方で，日本国内の出版者の DOI リンクや，Data Cite によるデータセット，オープンアクセスメガジャーナル「PLoS」の DOI リンクが含まれていることは，日本語版 Wikipedia が多様な学術情報と利用者を繋ぐ可能性を示している．JaLC をはじめ，DOI の付与対象を拡大する動きがあることから，それに伴って日本語版 Wikipedia における DOI リンクの件数やコンテンツ種別に変化が生じる可能性が考えられる．

今後の研究課題として，CrossRef 以外の RA，特に日本国内で刊行されるコンテンツを扱っている JaLC の DOI について詳細な分析を行うほか，件数だけではなく，実際の利用者がどのように DOI を利用しているかを明らかにする．これらの分析を通じて，DOI がどのような場所で，どのような学術情報を繋いでいるのかについて，より詳細に明らかにすることを目指す．

また，分析結果に見られた「存在しない DOI リンク」や「無効な DOI リンク」について，原因が DOI の記述の誤りであるか，DOI のシステム側でのリンク切れによるものであるかについても明らかにする．

参考文献

[1] ニールセン株式会社. “TOPS OF 2014: DIGITAL IN JAPAN ～ ニールセン 2014 年日本のインターネットサービス利用者数ランキングを発表 ～”. ニールセン株式会社. http://www.netratings.co.jp/news_release/2014/12/Newsrelease20141216.html,(参照 2015-04-14).

[2] Geoffrey Bilder. “Many Metrics. Such Data. Wow.”. CrossTech. 2014-02-24. http://crosstech.crossref.org/2014/02/many-metrics-such-data-wow.html,(参照 2015-04-14).

[3] Finn Arup Nielsen. Scientific citations in Wikipedia. First Monday. 2007, Vol.12, No.8, p.1-5. http://dx.doi.org/10.5210/fm.v12i8.1997, (参照 2015-04-14).

[4] 佐藤翔, 吉田光男, 安蒜孝政, 逸村裕. “日本語版 Wikipedia からの外部リンクの特徴とリンク切れの発生状況”. 第 19 回 (2011 年度) 研究報告会論文集. 香川, 2011-05-28/29. 情報知識学会, 2011, p.157-162. http://dx.doi.org/10.2964/jsik.21_06, (参照 2015-04-14).

[5] CrossRef. “crossref.org”. crossref.org. http://www.crossref.org/, (参照 2015-04-14).

[6] Japan Link Center. “ジャパンリンクセンター（JaLC）”. ジャパンリンクセンター（JaLC）. http://japanlinkcenter.org/, (参照 2015-04-14).

[7] Anna Tolwinska. “CrossRef Blog: Find the Registration Agency for any DOI”. CrossRef Blog. 2013-05-30. http://www.crossref.org/crweblog/2013/05/find_the_registration_agency_f.html, (参照 2015-04-14).

[8] “CrossRef/rest-api-doc”. CrossRef. https://api.crossref.org/, (参照 2015-04-14).

[9] “ノート:きんさんぎんさん”. Wikipedia. https://ja.wikipedia.org/wiki/%E3%83%8E%E3%83%BC%E3%83%88:%E3%81%8D%E3%82%93%E3%81%95%E3%82%93%E3%81%8E%E3%82%93%E3%81%95%E3%82%93, (参照 2015-04-14).

[10] THOMSON REUTERS. “Master Journal List - IP & Science”. INTELLECTUAL PROPERTY & SCIENCE. http://ip-science.thomsonreuters.com/mjl/, (参照 2015-04-14).

[11] THOMSON REUTERS. “JOURNAL LIST FOR ESSENTIAL SCIENCE INDICATORS”. SCIENCEWATCH. http://sciencewatch.com/info/journal-list, (参照 2015-04-14).