日本語版および英語版WikipediaにおけるDOIリンクの重複分析

著者	吉川 次郎, 佐藤 翔, 高久 雅生, 逸村 裕
著者別名	KIKKAWA Jiro, TAKAKU Masao, ITSUMURA Hiroshi
内容記述	第14回情報メディア学会研究大会 2015年6月27日 同志社大学 今出川キャンパス 良心館
雑誌名	第14回情報メディア学会研究大会発表資料
ページ	27-30
発行年	2015
URL	http://hdl.handle.net/2241/00125076

日本語版および英語版 Wikipedia における DOI リンクの重複分析

Duplicates of DOI Links between Japanese and English Wikipedia

吉川次郎[1]，佐藤翔[2]，高久雅生[3]，逸村裕[3]
Jiro KIKKAWA[1]，Sho SATO[2]，Masao TAKAKU[3]，Hiroshi ITSUMURA[3]
[1]筑波大学大学院図書館情報メディア研究科，
[2]同志社大学免許資格課程センター，
[3]筑波大学図書館情報メディア系
[1]Graduate School of Library, Information and Media Studies, University of Tsukuba,
[2]Center for License and Qualification, Doshisha University,
[3]Faculty of Library, Information and Media Science, University of Tsukuba

あらまし：日本語版 Wikipedia における DOI リンクが記述された経緯を明らかにするため、英語版 Wikipedia の翻訳項目の特定およびそこに含まれる DOI リンクの分析を行った。その結果、DOI リンクが記述されている日本語版の約 94%の項目が言語間リンクをもっていること、また、それらの項目間での共通の DOI リンクは、英語版の翻訳を通じて日本語版に記述されたものが大部分を占めることを示唆する結果が得られた。

キーワード：Wikipedia、Digital Object Identifier(DOI)、学術情報流通、Altmetrics

1．はじめに

学術情報流通の電子化を背景に、学術論文以外の情報源から学術情報が引用される機会が増えている。特に、ハイパーリンクを通じたウェブ上での学術情報の参照は、従来の学術論文からの引用に基づく評価指標に対して、代替的な評価指標(Altmetrics)として注目されつつある。さらに、オープンアクセスやオープンサイエンスをキーワードに、研究成果や研究データなどをウェブ上で公開する動きがあり、誰もがウェブ上で学術情報を容易に利用できる環境の構築が進んでいる。

ウェブ上の学術情報の参照に関する分析として、誰でも編集できるフリー百科事典であるWikipediaを対象とした研究が行われている。日本国外では、英語版Wikipedia(以下、英語版とする)に含まれる学術情報に着目し、英語版で多く参照されている学術論文とそのImpact Factorの値との関係性の分析を行ったNielsenの研究[1]がある。日本国内では、日本語版Wikipedia(以下、日本語版とする)について、学術論文の参照状況を分析した佐藤らの研究[2]がある。

学術情報を同定識別する仕組みとして、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るDOI(Digital Object Identifier)がある。DOIは、コンテンツのURLにリダイレクトするハイパーリンクとしての機能(以下、DOIリンクとする)をもつユニークな識別子であり、持続的なリンクを提供する点が特徴である。DOIリンクがどのような場所で、どのような学術情報を参照しているかについては、吉川らの日本語版におけるDOIリンクの分析[3]がある。

しかし、日本語版におけるDOIリンクについて、日本語版の項目作成時に独自に記述されたものであるか、他言語版の項目の翻訳を通じて記述されたものであるかなど、日本語版の項目におけるDOIリンクの記述の経緯については明らかにはなっていない。そこで、本研究では、このような経緯を明らかにすることを目的に、英語版と日本語版におけるDOIリンクの関係を明らかにする。

2. 対象と方法

2.1 分析対象

本研究では、2015年3月13日時点の日本語版と2015年3月4日時点の英語版における、外部リンク、項目情報、言語間リンクのダンプデータを用いる。DOIリンクは「dx.doi.org」または「doi.org」を含む外部リンクのうち、非DOIリンクを除外したうえで、百科事典の項目を意味する、名前空間が「0」である項目に含まれるDOIリンクとした。以下では、これらのDOIリンクを含む項目を分析対象とする。分析対象の異なり項目数は日本語版が9,135件、英語版が166,368件、のべDOIリンク数は日本語版が27,201件、英語版が1,473,728件であった。

2.2 分析方法

まず、日本語版と英語版のDOIリンクにおける重複状況について概観するため、DOIリンクの差集合および積集合を取得する。差集合は一方の言語版のみに記述されているDOIリンク、積集合は共通して記述されているDOIリンクである。

次に、英語版の項目の翻訳を通じて記述されたDOIリンクについて分析するため、分析対象の日本語版の項目について、(1)英語版の項目への言語間リンクが設定され、(2)共通のDOIリンクが10件以上あり、(3)編集履歴に英語版の翻訳である旨の記述を含む項目を特定する。

Wikipediaは言語版ごとに使用言語や記事編集者が異なるほか、各言語版で合意形成のうえで編集が行われるものではないため、同一主題の項目が言語版ごとに独立して存在している。言語間リンクは各言語版の同一主題へのリンクである。例えば、日本語版の「鳥類」は英語版の「Bird」への言語間リンクが設定されている。したがって、言語間リンクを用いることで、同一主題の特定が可能である。分析対象の日本語版の項目9,135件のうち言語間リンクが設定されている項目は8,574件(約94%)であった。これらの項目について、日本語版の項目と英語版の項目それぞれに共通するDOIリンクを集計した。さらに、共通のDOIリンクを10件以上含む日本語版の項目について編集履歴の最古10件と最新500件を確認し、編集履歴のコメント文に英語版の翻訳である旨の記述が含まれている場合は翻訳記事とみなし、その割合を算出する。

3. 分析結果

3.1 異なりDOIリンク数からみた日本語版と英語版の重複状況

異なりDOIリンク単位での日本語版と英語版の重複状況を示した表1の結果から、日本語版は英語版に比べ、全体での異なりDOIリンク数、非共通のDOIリンク数ともに少ないと言える。日本語版のみに記述されている異なりDOIリンク数は4,900であり、日本語版の約20%を占める。英語版と共通するDOIリンク数は19,171であり、日本語版の約80%に相当する一方で、英語版からみると約4%であり、英語版は日本語版が参照していないDOIリンクを多く含んでいる。

表 1 ： 異なり DOI リンク数からみた日本語版と英語版の重複状況

条件/言語版	日本語版	割合(%)	英語版	割合(%)
非共通の DOI リンク数(差集合)	4,900	20.36	500,209	96.31
共通の DOI リンク数(積集合)	19,171	79.64	19,171	3.69
合計	**24,071**	**100**	**519,380**	**100**

3.2 言語間リンクの有無からみた日本語版の項目とDOIリンク

言語間リンクの有無とのべDOIリンク数の関係を示した表2の結果から、日本語版におけるDOIリンクについて、言語間リンクが設定されている項目に含まれているものが、のべDOIリンク数の約94%を占めていることが分かる。さらに、この約94%のDOIリンクについて、英語版の同一主題項目におけるDOIリンクとの重複の結果を示した表3から、共通のDOIリンクを

含む項目が約86%を占めることが分かる。

共通DOIリンクを含む項目について、共通のDOIリンクを10件以上含む項目は、異なり項目数の約4%であるものの、のべDOIリンク数の約25%を占める。また、共通のDOIリンクを10件以上含む334件について、編集履歴の分析から英語版の項目を翻訳した旨の記述が確認された項目は323件であり、約97%に相当する。これらの項目に記述されているDOIリンクの多くは英語版の翻訳によるものと考えられる。一方で、共通のDOIリンクを1件も含まない項目は、異なり項目数の約22%を占めるものの、のべDOIリンク数でみると約14%にとどまっている。

表 2 : 言語間リンクの有無からみた日本語版の項目と DOI リンク

条件/項目	異なり項目数	割合(%)	のべ DOI リンク数	割合(%)
言語間リンクあり	**8,574**	**93.86**	**25,858**	**95.06**
言語間リンクなし	561	6.14	1,343	4.94
合計	**9,135**	**100**	**27,201**	**100**

表 3 : 言語間リンクが設定された項目と共通する DOI リンク

条件/項目	異なり項目数	割合(%)	のべ DOI リンク数	割合(%)
共通 DOI あり(10 件以上)	**334**	**3.90**	**6,707**	**25.94**
共通 DOI あり(10 件未満)	6,374	74.34	15,587	60.28
共通 DOI なし	1,866	21.76	3,564	13.78
合計	**8,574**	**100**	**25,858**	**100**

3.3 日本語版におけるDOIリンク数の多い項目

日本語版の項目について、DOIリンク数の多い上位10件を表4に示した。網掛箇所は日本語版の項目名およびDOIリンク数、その右側は日本語版の項目に対応する英語版の項目名とDOIリンク数である。共通DOIは、日本語版と英語版に共通するDOIリンク数を示している。上位10件の項目は、いずれも英語版の翻訳項目であることがわかる。また、共通するDOIリンク数から、英語版の翻訳によってDOIリンクが記述されていると考えられる。

表 4 : 日本語版における DOI リンク件数の多い項目(上位 10 件)

順位	項目名(日本語版)	DOI リンク	項目名(英語版)	DOI リンク	共通 DOI	翻訳
1	抗酸化物質	**165**	Antioxidant	392	142	翻訳
2	リーリン	**121**	Reelin	350	120	翻訳
3	鳥類	**119**	Bird	276	114	翻訳
4	ベンゾジアゼピン…	**85**	Benzodiazepine…	186	77	翻訳
5	免疫系	**84**	Immune system	196	80	翻訳
6	ポリアデニル化	**78**	Polyadenylation	162	78	翻訳
7	抗精神病薬	**68**	Antipsychotic	214	29	翻訳
8	抗うつ薬	**67**	Antidepressant	236	20	翻訳
9	エピジェネティクス	**62**	Epigenetics	234	29	翻訳
10	植物の進化	**58**	Evolution of plants	212	56	翻訳

3.4 言語間リンクが設定されている項目のうち、共通のDOIリンクを含まない項目の特徴

言語間リンクが設定されている項目のうち、共通のDOIリンクを含まない項目について、DOIリンク数の多い上位10件を表5に示した。「ロットレリン」は英語版の項目のReferencesと日本語版の項目の脚注が似通っており、英語版では書誌事項の記述にDOIリンクが使用されていないのに対して、日本語版の項目ではDOIリンクの書式に修正されている。「放射化学」についても「ロットレリン」と同様である。「セレブロシドスルファターゼ」は言語間リンクが誤って設定されている。これらの項目に含まれるDOIリンクは、英語版の項目から流入したものであることが疑われる。それら以外は、日本の人物に関する項目が含まれていることから、

英語版の翻訳ではなく日本語版由来のDOIリンクであると考えられる。

表 5: 言語間リンクが設定されており、共通の DOI リンクをもたない項目

順位	項目名	DOI リンク	翻訳
1	ロットレリン	17	翻訳
1	セレブロシドスルファターゼ	17	翻訳
3	藤田昌久	15	非翻訳
3	森政弘	15	非翻訳
3	今泉吉典	15	非翻訳
6	放射化学	14	翻訳
6	ジェームズ・ハミルトン_(計量経済学者)	14	翻訳
8	疲労骨折	13	非翻訳
8	出版バイアス	13	翻訳
8	巨人の肩の上	13	翻訳

4. 考察と今後の課題

本研究では、日本語版におけるDOIリンクが記述された経緯について明らかにするために、DOIリンクを含む日本語版の項目について、翻訳項目の特定および分析を行った。

分析の結果、DOIリンクが記述されている日本語版のほとんどの項目について、英語版への言語間リンクが設定されていることが分かった。さらに、言語間リンクが設定されている項目間での共通のDOIリンクの集計および当該記事の編集履歴の分析から、英語版の翻訳を通じた日本語版の項目へのDOIリンクの流入が、日本語版におけるDOIリンクの大部分を占めることを示唆する結果が得られた。したがって、日本語版におけるDOIリンクの分析では、他言語版からの翻訳を通じて記述されたDOIリンクを考慮する必要がある。また、この点については、DOIリンクに限らず、PubMedリンクなどについても他言語版の翻訳を通じたリンクの流入が考えられるため、Wikipedia上での学術情報の参照状況を分析するにあたっては各言語版由来の学術情報の参照と他言語版の翻訳を通じた学術情報の参照を区別する必要があると言える。

DOIリンクについてはAltmetricsスコアの算出に使用されていることから、もし、Altmetricsスコアの収集対象が多言語版のWikipediaに拡大した場合、英語版でのDOIリンクの参照が多重集計されることが懸念される。

今後の課題として、日本語版での学術情報の参照の特徴や性質を明らかにするために、他言語版からの翻訳などを通じて流入したDOIリンクを除外し、その特徴や性質の分析を行うことを検討する。

参考文献

[1] 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).

[2] 佐藤翔, 吉田光男, 逸村裕. Wikipedia 日本語版からの学術論文の引用状況. 2013 年日本図書館情報学会春季研究集会発表論文集. 茨城, 2013-05-25, p.81-84.

[3] 吉川次郎, 高久雅生, 逸村裕. 日本語版WikipediaにおけるDOIリンクの予備的分析. 第23回(2015年度)情報知識学会年次大会. 東京, 2015-05-23/24. 情報知識学会誌. 2015, Vol.25, No.2, p.160-165.