Wikipedia上の学術情報のLOD化に向けた予備的分析

Preliminary Analyses for Extracting Scholarly Information on Wikipedia as Linked Open Data

吉川次郎[1*]　高久雅生[2]　加藤文彦[3]
大向一輝[3]　武田英明[3]
Jiro KIKKAWA[1]　Masao TAKAKU[2]　Fumihiro KATO[3]
Ikki OHMUKAI[3]　Hideaki TAKEDA[3]

[1] 筑波大学大学院 図書館情報メディア研究科
[1] Graduate School of Library, Information and Media Studies, University of Tsukuba
[2] 筑波大学 図書館情報メディア系
[2] Faculty of Library, Information and Media Science, University of Tsukuba
[3] 国立情報学研究所
[3] National Institute of Informatics

Abstract: In this paper, the authors have attempted to build a dataset of scholarly information on Wikipedia as Linked Open Data. As a data model, we refer to "DBpedia citations & references challenge", hereafter DBpedia method. We extracted scholarly information from external links in Japanese Wikipedia. As of March 2016, there were 35,266 DOI links and 10,047 CiNii permalinks that identified with NAID in main namespace pages. As a result of comparison between DBpedia method and proposed method, DBpedia and proposed method extracted common DOI Links at high rate. DBpedia method extracted fewer CiNii permalinks than proposed method.

1　はじめに

学術情報流通の電子化により、Webを通じて誰もが学術情報を即時かつ容易に入手可能な環境が構築されている。このような環境において、Wikipediaをはじめとするオープンな Webが、Webの利用者と学術情報との橋渡しをするような役割を果たすと考え、これまでに筆者らは Wikipediaにおける「DOI(Digital Object Identifier、デジタルオブジェクト識別子)」を対象に、(1) 日本語版における学術情報の参照状況 [1]、(2) 学術情報の参照における言語版間での重複状況 [2]、(3)JaLC(Japan Link Center) DOI のアクセスログ分析を通じた参照元の分析 [3] に取り組んできた。結果、(1)2015 年 3 月時点の日本語版 Wikipedia において DOI を用いて 3 万件弱の学術情報の参照が行われている、(2) 英語版 Wikipedia の翻訳を通じて日本語版 Wikipedia に記述された学術情報の参照が見られる、(3)JaLC DOI のアクセス元のうち、6 番目にアクセスの多い参照元が日本語版 Wikipedia であることがそれぞれ明らかになった。

近年、Linked Open Data(以下、「LOD」) の技術を用いたデータセットの構築と活用が盛んである。たとえば、LOD のデータセット同士のネットワークを図示した 2014 年版の LOD cloud diagram [4] において、学術情報に関連するものとして、Wikipedia の情報を LOD 化したデータセットである DBpedia が中央に位置し、出版系分野として DOI が登場する。以上から、Wikipedia 上の DOI を用いた学術情報の参照を LOD 化することで、DBpediaをはじめとする他のデータセットと連携させることが可能であると考えられる。

本研究に関連する先行事例として、DBpediaによる、Wikipedia上の外部リンクを LOD 化するプロジェクトである「DBpedia citations & references challenge [5]」が挙げられる。同プロジェクトは試行段階であり、英語版 Wikipediaを用いたデータセットを公開したうえで、他言語版への拡張や、関連サービスとの連携などのアイデアを募っている。また、データ抽出およびデータセット構築のためのツールは、Wikipediaから構造化データを抽出するための DBpedia Information Extraction

*連絡先：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: jiro@slis.tsukuba.ac.jp

Framework [6] の一部として公開されている。

本研究では、日本語版 Wikipedia における学術情報の参照を LOD 化するための予備的な分析を行う。具体的には、日本語版 Wikipedia を対象に、(1) どのような学術情報がどれくらい参照されているかに関する調査、(2) 学術情報の参照記述の精緻化に向けた検討を行う。データモデルに関しては、同プロジェクトの設計を参照する。データセットの構築において、同プロジェクトで公開されているツールを日本語版 Wikipedia に適用したうえで学術情報の参照のみを抽出する方法があり得る。しかし、この方法によって日本語版 Wikipedia 上の DOI を用いた学術情報の参照がどの程度の精度で抽出可能であるか、日本語版 Wikipedia に適用可能であるか否かは不明である。これらの理由から、本研究では、DBpedia citations & references challenge での方法論 (以下、「DBpedia 手法」) と、筆者らが提案する手法 (以下、「提案手法」) の比較による検討を行う。

以上の分析から得られた知見は、DBpedia citations & references challenge プロジェクトに対するフィードバックとしての価値があるほか、日本語版 Wikipedia の学術情報の参照を LOD 化するうえでの検討材料としての価値があると考えられる。また、応用として、DBpedia への反映、各言語版の同一主題ページで参照されている学術情報の統合や提示、他のデータセットとの連携などが挙げられる。

2 関連研究

本研究における関連研究として、(1)Wikipedia 上の外部リンクの生存状況、(2)Wikipedia 上の学術情報の参照状況、(3)Wikipedia を通じた学術情報へのアクセス状況の分析について述べる。

2.1 Wikipedia 上の外部リンクの生存状況

Tzekou ら [7] は、2009 年 10 月時点の英語版 Wikipedia を対象に外部リンクの生存状況の分析を行った。ページごとのリンク切れの分布を調査した結果から、外部リンクの約 18%でリンク切れが生じていることを明らかにした。ただし、リンク切れの多くは少数のページに記述された外部リンクに集中していること、約 77%のページにおける外部リンクに関して、リンク切れがまったく生じていないことから、Wikipedia における外部リンクの多くは到達可能なものであると指摘している。

佐藤ら [8] は、日本語版 Wikipedia における外部リンクの特徴およびリンク切れの発生状況に着目した分析を行った。2011 年 4 月 20 日時点での日本語版 Wikipedia を対象とした分析から、(1) 日本語版の外部リンクの 11%程度でアクセスに障害があること、(2) edu, co.jp, go.jp ドメインにおいてアクセス障害が多いこと、(3) 新聞社が運営するニュースサイトで特にアクセス障害が多いことを指摘している。

2.2 Wikipedia 上の学術情報の参照状況

Nielsen [9] は、2007 年 4 月時点の英語版 Wikipedia の外部リンクに含まれる学術情報に着目し、英語版 Wikipedia において多く参照されている学術論文と、当該論文の Journal Citation Reports における Impact Factor の値との関係の分析を行った。分析結果から、参照されている学術情報について、Nature、Science などのジャーナルが多いこと、天文学分野のジャーナルが多いこと、必ずしも Impact Factor の値が高いジャーナルが多いわけではないことを指摘している。

Lin ら [10] は、総記事数の多い 25 の言語版を対象に、DOI を用いて PLOS のコンテンツの参照状況の分析を行った。結果、2014 年 3 月時点での PLOS のすべてのコンテンツのうち、4%が Wikipedia 上で 1 回以上参照されていること、それらのうち 47%が英語版以外の言語版から参照されていることが明らかになった。また、「PLOS コンテンツを引用しているページの多さ」と「言語版におけるアクティブなユーザーの多さ」に強い正の相関が見られることを指摘している。

佐藤ら [11] は 2012 年 12 月時点の日本語版 Wikipedia を対象に、学術論文の引用状況を分析している。分析対象は「PubMed」、「CiNii Articles」、「機関リポジトリ」である。分析結果から、PubMed に関しては英語版 Wikipedia に比べて引用が少ないこと、CiNii Articles と機関リポジトリについては、それぞれのサービスの収録論文数の規模を考慮すると引用件数としては少数に留まっていることを指摘している。

学術情報の参照状況について、特に DOI に着目したものとして、吉川らは、2015 年 3 月時点の日本語版 Wikipedia を対象とした分析を行い、97%が Crossref DOI、2%が JaLC DOI であること、日本国外の大手出版者のコンテンツの参照が多いことを明らかにした [1]。さらに、これらの参照について、言語間リンクを用いて日本語版と英語版の同一主題ページを調査し、英語版の翻訳を通じて日本語版に流入した記述が大部分を占めることを示唆する結果が得られたとしている [2]。Mietchen らは、DOI を含む各識別子の記述件数について、2015 年 6 月 22 日時点の英語版とオランダ語版を対象とした分析から、両言語版ともに ISBN の件数が最も多く、次いで DOI が多いという結果のほか、時系列での件数の推移を示している [12, 13]。

その他、Crossref Labs や Wikimedia Tool Labs によって、DOI を通じた学術情報の参照を可視化するサービスが開発されている。「Wikipedia Cite-o-Meter [14]」

は、日本語版や英語版など 100 種類の言語版に記述されている DOI 名について、Prefix 単位での参照状況を表示するサービスである。たとえば、PLoS(Public Library of Science) の Prefix である「10.1371」を含む DOI 名が日本語版や英語版にどれだけ記述されているかを表示することが可能である。「DOI Chronograph [15]」は、Crossref DOI について、DOI 名、参照元ドメイン名、参照元サブドメイン名ごとに、アクセス数を表示するサービスである。また、時系列形式でのアクセス数の推移を表示することができる。

2.3 Wikipedia を通じた学術情報へのアクセス状況の分析

世界最大規模の DOI 登録機関である Crossref は、Crossref DOI の参照元およびアクセス状況について、アクセスログを用いた分析結果の報告を行っている。2015 年時点での報告 [16] では、Crossref DOI の参照元の上位 4 件は学術文献データベースであり、それぞれ、Web of Science、Serials Solutions、ScienceDirect、Scopus である。それらに次いで、5 番目に大きな参照元が Wikipedia である。Wikipedia からのアクセスを言語版ごとに見たときの上位 10 件は、それぞれ、英語版、英語版 (モバイル)、ドイツ語版、日本語版、スペイン語版、フランス語版、ロシア語版、中国語版、イタリア語版、ポルトガル語版である。

日本国内唯一の DOI 登録機関である JaLC によって登録された JaLC DOI の参照元およびアクセス状況については、吉川ら [3] によるアクセスログの分析事例がある。2014 年 4 月から 2015 年 9 月時点におけるアクセスを対象に、完全修飾ドメイン名ごとの集計を行った結果から、6 番目にアクセスの多い参照元が日本語版 Wikipedia であることが明らかになった。

3 DOI とは

DOI(Digital Object Identifier、デジタルオブジェクト識別子) とは、解決可能、持続可能、相互運用可能なリンクを提供するための基盤である。2016 年 8 月現在、DOI 名の登録件数は約 1 億 3,000 万 [17] である。

DOI 名は、「10.」からはじまる文字列である Prefix、「/」、Suffix で構成される。DOI 名を “http://doi.org/”(または “http://dx.doi.org/”) の後ろに加えることで、当該コンテンツの URI へのリダイレクトを行うハイパーリンクとして機能する。本研究では、この URI を通じたハイパーリンクのことを「DOI リンク」と呼ぶ。たとえば、DOI 名「10.1241/johokanri.56.414」の場合、Prefix は「10.1241」、Suffix は「johokanri.56.414」、DOI リンクは「http://doi.org/10.1241/johokanri.56.414」(または「http://dx.doi.org/10.1241/johokanri.56.414」) である。

Prefix は学協会や出版社など、DOI 登録者ごとに割り与えられる。DOI 登録者は自身のコンテンツに Suffix を割り与え、DOI 登録機関 (Registration Agency、以下、「RA」) を通じて DOI 名の登録を行う。2016 年 8 月現在、RA は全 10 機関である。それらのうち、ジャーナル記事、書籍、データセットなどの学術情報を扱う RA として、Crossref、JaLC、ISTIC、DataCite がある。

4 DBpedia 手法

本章では、DBpedia 手法の抽出方法とデータモデルについて述べる。

DBpedia 手法は、Wikipedia のダンプデータのうち、会話ページおよび利用者ページを除く各ページの本文内容をウィキテキスト形式で格納した「pages-articles.xml」を用いて、標準名前空間における出典テンプレートの抽出を行ったうえで、当該 Template の記述内容から、(1) リソースの URI、(2) 書誌要素、(3) 書誌要素の値の取得を行う。抽出対象となる出典テンプレートの記述例を図 1、抽出結果を図 2 および図 3 に示す。

図 1 は、日本語版 Wikipedia の「オープンアクセス」における Cite journal テンプレートの例である。「タイトル (title)」、「著者 (author)」、「誌名 (journal)」、「DOI 名」などの書誌要素と、その値が記述されている。

```
{{Cite journal|title=オープンアクセスの広がりと
現在の争点|author=佐藤翔|journal=情報管理
|issue=7|volume=56|year=2013|publisher=
[[科学技術振興機構]]|pages=414-424
|doi=10.1241/johokanri.56.414|ref=佐藤 2013}}
```

図 1: 日本語版 Wikipedia の「オープンアクセス」における Cite journal テンプレートを用いた記述例

DBpedia 手法は、出典テンプレートを用いて 2 つのデータセットの構築を行う。それぞれ、図 2 に示す「citation-links.ttl」、図 3 に示す「citation-data.ttl」である。citation-links.ttl は、リソースを「参照されている外部リンクの URI」、プロパティを「http://ja.dbpedia.org/property/isCitedBy」、プロパティの値を「参照元の Wikipedia のページ」とするトリプルである。citation-data.ttl は、リソースを「参照されている外部リンクの URI」、プロパティを「title(タイトル)」や「author(著者)」などの書誌要素、プロパティの値を実際の書誌要素の値とするトリプルである。すなわち、citation-links.ttl は「参照されている外部リンク」と「参照元の Wikipedia のページ」の対応関係を示し、

citation-data.ttl は「参照されている外部リンク」と「外部リンクの書誌要素およびその値」の対応関係を示す。

DBpedia 手法の特徴として、テンプレートに DOI 名が記述されている場合は「http://doi.org/」の後ろに DOI 名を追加することで URI を生成する。同一ページ内の出典テンプレートのパラメーターにおいて、「doi」の値に DOI 名が記述されており、尚且つ、「url」に「http://dx.doi.org/」のような形式で DOI リンクが記述されている場合は、これらの記述それぞれが 1 行ずつ記録される。ISBN が記述されている場合は「http://books.google.com/books?vid=ISBN」の後ろに ISBN を追加した URI を生成する。たとえば、日本語版の「Ruby」のページにおいて ISBN「978-4-8222-3431-7」が出典テンプレート内に記述されている場合、「http://books.google.com/books?vid=ISBN978-4-8222-3431-7」となる。以上は一例であるが、これらの処理が行われることから、実際の外部リンクの URI と DBpedia 手法でのリソースの URI は、必ずしも同一とは限らない。

```
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/isCitedBy>
<http://ja.dbpedia.org/resource/オープンアクセス>
.
```

図 2: citation-links.ttl の例

```
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/title> “オープンアクセスの広がりと現在の争点”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/author> “佐藤翔”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/journal> “情報管理”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/issue> “7”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/volume> “56”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/year> “2013”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/publisher>
<http://ja.dbpedia.org/resource/科学技術振興機構>
.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/pages> “414-424”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/doi>
“10.1241/johokanri.56.414”.
<http://doi.org/10.1241/johokanri.56.414>
<http://ja.dbpedia.org/property/ref> “佐藤 2013”.
```

図 3: citation-data.ttl の例

5 対象と方法

2016 年 3 月 5 日時点の日本語版 Wikipedia における学術情報の参照状況の分析および LOD 化に向けた検討として、(1) 参照の多い外部リンクおよび学術情報の特定、(2) 学術情報の抽出状況の分析、(3) 書誌要素の抽出状況の分析を行う。

5.1 参照の多い外部リンクおよび学術情報の特定

DBpedia 手法と提案手法のそれぞれを用いて生成したデータセットを用いて、日本語版 Wikipedia において参照の多い外部リンクおよび学術情報の特定を行う。

筆者らのこれまでの分析から、DOI リンクについては、日本語版 Wikipedia に 3 万件程度記述されていることが明らかになっているが、その他の学術情報については必ずしも参照状況は明らかではない。したがって、まず、外部リンクをすべて抽出したうえで FQDN 単位での集計を行い、参照が多く、尚且つ、Web API などを通じて書誌要素が取得可能な学術情報を特定する。

分析の結果、DOI リンクに加えて CiNii が多く参照されていることが明らかになったため、DOI リンクと CiNii のコンテンツを分析対象とする。

5.2 学術情報の抽出状況の分析

DOI リンクおよび CiNii の参照について、DBpedia 手法と提案手法の抽出結果の比較を行う。すなわち、DBpedia 手法における citation-links.ttl と提案手法による抽出結果を用いて、異なり URI および異なり参照元ページタイトルの重複状況について分析を行う。

もし、それぞれの重複状況の分析結果から、両者の重複率が高いことが明らかになった場合、DBpedia 手法による方法論を日本語版 Wikipedia に適用することで学術情報の参照を LOD 化することが可能であると言える。その一方で、もし、DBpedia 手法と提案手法の間に大きな差異が見られた場合は、DBpedia 手法の実装の見直しや改善などについて検討を行う余地があるほか、各言語版に対して一律に適用することは難しいことが指摘できる。

5.3 書誌要素の抽出状況の分析

DBpedia 手法における citation-data.ttl を対象に、書誌要素の抽出状況の検討を行う。DBpedia 手法は出典テンプレートの情報を抽出することで書誌要素およびその値の取得を行うため、出典テンプレート自体に情報が記述されていない場合は、プロパティの値を抽出

することができない。したがって、この分析は、出典テンプレート自体にどのような書誌要素が記述されているかに関する検討を兼ねたものである。

もし、特定の書誌要素の抽出率が低い場合は、データセット構築の方法論として、Web API などの連携を通じた書誌要素のカバレッジ向上などの検討が課題として挙げられる。なお、テンプレートの情報に含まれる書誌要素の値自体の内容や正確性に関しても検討の必要があるが、その点については今後の課題とする。

5.4 提案手法

2016 年 3 月 5 日時点の日本語版 Wikipedia のダンプデータのうち、外部リンクを格納した「externallinks.sql」、インターウィキリンク (ウィキ間リンク) を格納した「iwlinks.sql」、ページ情報を格納した「pages.sql」を使用し、百科事典ページを意味する標準名前空間において参照されている外部リンクの抽出を行う。

抽出方法は、まず、pages.sql と externallinks.sql を用いて「page_id(ページの id)」、「page_title(ページのタイトル)」、「page_namespace(ページの属する名前空間)」と「el_to(外部リンクの URI)」を取得する。次に、DOI リンクについては externallinks.sql 以外に、iwlinks.sql に格納されるケースがあるため、pages.sql と iwlinks.sql を用いて「page_id」、「page_title」、「page_namespace」、「iwl_prefix(インターウィキリンクの種別)」、「iwl_title(インターウィキリンクの値)」を取得する。以上の方法で抽出した内容を統合したものを標準名前空間における外部リンクおよび参照元ページのデータとして用いる。

6 分析結果と考察

6.1 参照の多い外部リンクおよび学術情報

参照の多い外部リンクおよび学術情報について、DBpedia 手法および提案手法での FQDN ごとの集計結果を表 1、表 2 に示す。

表 1 において最も件数の多い「citation.dbpedia.org」は、出典テンプレートが記述されているもののリソースとなる URI が存在しない場合に生成されるものである。つまり、出典テンプレート自体は記述されているものの、「url」、「doi」、「isbn」などの値が記述されていない場合に生成される URI であるため、実際には外部リンクではない。「books.google.com」は、表 1 では 2 位、表 2 では 20 位であり、件数に大きな差がある。この点は、4 章で述べたように、出典テンプレートに ISBN の値が記述されている場合に生成される URI の FQDN であるためと考えられる。これら 2 項目を除き、DBpedia 手法よりも提案手法のほうが抽出件数が多い。

参照の多い学術情報としては DOI リンクがあり、表 1 の 3 位である「doi.org」、表 2 の 9 位である「dx.doi.org」が該当する。表 1 と表 2 において DOI リンクの FQDN が異なるのは、4 章で述べた URI の生成方法の差異に起因するものである。それぞれの手法における DOI リンクの抽出件数を表 3、表 4 に示す。

表 3、表 4 の結果から、DOI リンクについては、DBpedia 手法と提案手法での件数の差が他の外部リンクに比べて小さく、DBpedia 手法における抽出率が高いと言える。表 3 の「エラー (Bad URI)」は、DBpedia 手法において不正な記号を含む URI として処理された項目である。具体的には、「ブラケット ([])」、「半角スペース」、「バックスラッシュ (\)」、「不等号 (<)」、「番号記号 (#)」が含まれる場合である。これらのうち、半角スペースとバックスラッシュを除く記号については、DOI 名の Suffix に使用されることのある記号である。以上から、参照の多い学術情報のうち DOI リンクに関しては、件数で見た場合の抽出率は高いと言えるものの、実際の URI として正確な抽出が行われているかに関しては疑問が残るため、次節にて追加分析を行う。

学術情報の参照として、DOI リンクのほかに、表 2 の 13 位に CiNii がある。以下、CiNii のコンテンツである、「ci.nii.ac.jp」ではじまる URI を「CiNii URI」と呼ぶ。DBpedia 手法での CiNii URI は 50 位 (1,651 件) であり、提案手法での CiNii URI の抽出件数と比較した場合、8%相当である。CiNii URI は DOI リンクと同様に、Web API を通じた書誌要素の取得が可能であるため、以降の分析では DOI リンクおよび CiNii URI を分析対象とする。

6.2 学術情報の抽出状況の分析結果

6.2.1 DOI リンクの抽出状況

DBpedia 手法における citation-links.ttl と提案手法における DOI リンクについて、異なり URI および異なり参照元ページタイトルの重複状況の分析結果を示す。異なり URI については、前節での FQDN ごとの集計結果を踏まえ、異なり DOI 名を用いた分析を行うこととし、表 3 の「エラー (Bad URI)」を除く 29,000 件と提案手法における 35,266 件を用いる。ページタイトルの異なり数は、DBpedia 手法が 9,388 件、提案手法が 11,542 件である。これらを用いて、共通要素 (積集合) および一方のみから参照が行われている要素 (差集合) の取得による分析を行う。

まず、異なり DOI 名の重複状況について、共通の DOI 名は 25,520 件、DBpedia 手法のみに登場する DOI 名は 219 件、提案手法のみに登場する DOI 名は 5,394 件であった。ただし、Turtle において URI を囲む記号として使用される「<」や「>」が DOI 名に含まれている場合に、抽出結果に差異が生じているケースが

表 1: DBpedia 手法の citation-links.ttl におけるリソースの FQDN ごとの集計結果 (上位 20 件、n=1,105,485)

順位	FQDN	概要	件数
1	citation.dbpedia.org	–	137,747
2	books.google.com	Google Books	121,949
3	doi.org	Digital Object Identifier System	28,593
4	ameblo.jp	アメーバブログ	11,558
5	natalie.mu	ナタリー	10,306
6	www.oricon.co.jp	ORICON STYLE	9,927
7	www.sponichi.co.jp	スポニチ Sponichi Annex	7,819
8	www.ncbi.nlm.nih.gov	National Center for Biotechnology Information	6,427
9	sankei.jp.msn.com	MSN 産経ニュース	6,346
10	www.nikkansports.com	日刊スポーツ	5,975
11	www.asahi.com	朝日新聞デジタル	5,741
12	www.yomiuri.co.jp	読売新聞 (YOMIURI ONLINE)	5,310
13	www.worldcat.org	OCLC WorldCat	5,010
14	mainichi.jp	毎日新聞のニュース・情報サイト	4,858
15	dawinci.istat.it	Dawinci home page	4,311
16	www.47news.jp	47NEWS(よんななニュース)	4,222
17	www.cinematoday.jp	シネマトゥデイ	4,010
18	www.wrestling-titles.com	Pro-Wrestling Title Histories: championshi	3,945
19	www.census.gov	United States Census Bureau	3,785
20	www.baseball-reference.com	Baseball-Reference.com	3,721

表 2: 提案手法での抽出結果における FQDN ごとの集計結果 (上位 20 件、n=4,341,678)

順位	FQDN	概要	件数
1	tools.wmflabs.org	Wikimedia Tool Labs	353,277
2	commons.wikimedia.org	Wikimedia Commons	212,027
3	web.archive.org	Internet Archive: Wayback Machine	57,610
4	twitter.com	Twitter	47,600
5	www.worldcat.org	OCLC WorldCat	40,228
6	ameblo.jp	アメーバブログ	39,292
7	www.imdb.com	IMDb - Movies, TV and Celebrities	36,898
8	www.youtube.com	YouTube	34,878
9	dx.doi.org	DOI	32,095
10	www.ncbi.nlm.nih.gov	National Center for Biotechnology Information	29,133
11	www.allcinema.net	映画データベース - allcinema	21,236
12	www.sponichi.co.jp	スポニチ Sponichi Annex	20,511
13	ci.nii.ac.jp	CiNii	20,418
14	viaf.org	VIAF(バーチャル国際典拠ファイル)	19,227
15	dl.ndl.go.jp	国立国会図書館デジタルコレクション	18,878
16	sankei.jp.msn.com	MSN 産経ニュース	15,351
17	www.nikkansports.com	日刊スポーツ	14,770
18	www.oricon.co.jp	ORICON STYLE	14,537
19	www.facebook.com	Facebook	14,497
20	books.google.com	Google Books	14,441

表 3: DBpedia 手法による DOI リンクの抽出件数

FQDN	抽出元	件数
doi.org	citation-links.ttl	28,593
dx.doi.org	citation-links.ttl	407
エラー (Bad URI)	citation-links.ttl	113
合計	–	**29,113**

表 4: 提案手法による DOI リンクの抽出件数

FQDN	抽出元	件数
dx.doi.org	externallinks.sql	32,095
dx.doi.org	iwlinks.sql	2,175
doi.org	externallinks.sql	995
www.doi.org	externallinks.sql	1
合計	–	**35,266**

見られた。たとえば、DBpedia 手法と提案手法において、同一 DOI 名に対する抽出結果が、それぞれ、「10.1002/(sici)1096-8628(19981116)80:33.0.co;2-g」と「10.1002/(sici)1096-8628(19981116)80:3<204::aid-ajmg4>3.0.co;2-g」であった場合が挙げられる。このとき、DOI 名および DOI リンクとして有効であるのは提案手法の値であり、DBpedia 手法の抽出結果については、参照されている学術情報にアクセスすることができないため、エスケープ処理を行う必要がある。なお、提案手法で取得した DOI 名について、「<」から「>」まで」の文字列を削除したうえで、再度、差集合を取得した結果、DBpedia 手法のみに登場する DOI 名は 219 件から 40 件に減少した。このことから、DBpedia 手法において、「<」や「>」の記号が含まれているために DOI 名および DOI リンクの正確な抽出が行われていないケースが複数存在することが分かった。

次に、参照元ページタイトルの重複状況については、共通のページタイトルが 9,338 件、DBpedia 手法のみに登場するページタイトルが 11 件、提案手法のみに登場するページタイトルが 2,203 件であった。11 件の詳細を調査した結果、実際には共通のページタイトルに含まれるものが 1 件、ページの移動に伴うタイトル変更によるものが 1 件、テンプレート記述におけるパラメーター指定に誤りがあると考えられるものが 3 件、英語版ページの翻訳を行った際に英語版のテンプレートをそのまま使用し、日本語版の当該テンプレートのパラメーターに整合していないことが原因と考えられるものが 3 件、パラメーターとして「doi_brokendate」が指定されているために DOI リンクが生成されていないものが 3 件であった。

以上の結果から、DBpedia 手法による URI の抽出処理に関して、正確な URI の抽出が行われていないものが含まれていること、実際の日本語版 Wikipedia 上では外部リンクとして現れないものが含まれていることが明らかになった。

6.2.2 CiNii URI の抽出状況

DBpedia 手法における citation-links.ttl における CiNii URI と、提案手法における CiNii URI について、URI および参照元ページタイトルの重複状況の分析結果を示す。CiNii URI は、DOI リンクとは異なり、論文などの学術情報本体を示す URI のほかに、検索ページや著者情報のページなどの URI が存在する。したがって、CiNii URI の抽出状況の検討においては、まず、どのようなコンテンツが参照されているかを調査し、LOD 化の対象範囲の検討を行う。

DBpedia 手法および提案手法における CiNii URI の内訳を表 5、表 6 にそれぞれ示す。CiNii URI の内訳については、「http://ci.nii.ac.jp/」以降の最初の「/」まで、「/」が含まれない場合は「?」までをそれぞれ抽出したうえで集計を行った。

表 5: DBpedia 手法における CiNii URI の内訳 (n=1,651)

順位	項目	件数
1	http://ci.nii.ac.jp/naid/	1,242
2	http://ci.nii.ac.jp/ncid/	219
3	http://ci.nii.ac.jp/els/	111
4	http://ci.nii.ac.jp/lognavi?	46
5	http://ci.nii.ac.jp/vol_issue/	8
5	http://ci.nii.ac.jp/search?	8
7	http://ci.nii.ac.jp/nrid/	4
8	http://ci.nii.ac.jp/organ/	3
8	http://ci.nii.ac.jp/author?	3
10	http://ci.nii.ac.jp/info/	2
10	http://ci.nii.ac.jp/Detail/	2
12	http://ci.nii.ac.jp/	1
12	http://ci.nii.ac.jp/author/	1
–	その他	1

表 5、表 6 の結果から、DBpedia 手法と提案手法のいずれにおいても最も件数が多いのは論文詳細ページであり、DBpedia 手法による論文詳細ページの抽出件数は提案手法の抽出件数の約 12%である。以下、CiNii URI については、論文詳細ページの「http://ci.nii.ac.jp/naid/」を対象に、書誌要素の抽出状況の分析を行う。

表 6: 提案手法における CiNii URI の内訳 (n=20,418)

順位	項目	件数
1	http://ci.nii.ac.jp/naid/	10,047
2	http://ci.nii.ac.jp/ncid/	3,878
3	http://ci.nii.ac.jp/author/	3,114
4	http://ci.nii.ac.jp/search?	1,273
5	http://ci.nii.ac.jp/els/	652
6	http://ci.nii.ac.jp/author?	514
7	http://ci.nii.ac.jp/nrid/	393
8	http://ci.nii.ac.jp/vol_issue/	206
9	http://ci.nii.ac.jp/lognavi?	193
10	http://ci.nii.ac.jp/cinii/	35
11	http://ci.nii.ac.jp/organ/	27
12	http://ci.nii.ac.jp/books/	24
13	http://ci.nii.ac.jp/Detail/	22
14	http://ci.nii.ac.jp/	16
15	http://ci.nii.ac.jp/search/	3
15	http://ci.nii.ac.jp/info/	3
15	http://ci.nii.ac.jp/d/	3
18	http://ci.nii.ac.jp/openurl/	1
18	http://ci.nii.ac.jp/keyword/	1
–	その他	13

6.3 書誌要素の抽出状況の分析結果

DBpedia 手法での citation-data.ttl における書誌要素の抽出状況について、外部リンク全体を対象とした集計結果を表 7、DOI リンクを対象とした集計結果を表 8、CiNii URI のうち論文詳細ページを対象とした集計結果を表 9 に示す。

citation-data.ttl は、リソースが「参照されている URI」、プロパティが「書誌要素」、プロパティの値が「実際の書誌要素」のトリプルである。ただし、同一リソースが複数回に渡って参照されている場合、これらを区別することなく、それぞれ 1 行ずつ保存している。したがって、書誌要素ごとの分析にあたっては、予め、リソースとプロパティのペアを抽出し、重複を除外したうえで、それぞれのリソースについて、書誌要素ごとの抽出率を算出した。DBpedia 手法におけるプロパティの値は、出典テンプレートの記述内容に依存しており、抽出率が高い項目は、パラメーターとして記述されていることが多い項目、抽出率が低い項目は記述されていることが少ない項目を意味する。

表 7 の結果から、DBpedia 手法によって抽出されたリソース全体において、書誌要素の抽出率が 50%以上の項目は、「title(タイトル)」、「url」、「accessdate(参照日)」、「publisher(出版者)」である。この結果から、DBpedia 手法での方法論を用いることで、書誌要素のうち、これらの 4 項目については取得可能である場合が比較的多いと言える。ただし、これらはあくまで件数のみを見た場合の結果であり、プロパティの値の正確性については、本研究での分析の対象範囲外である。

表 8 の結果から、学術情報の参照記述における書誌要素に関して、DOI リンクは、「title」、「doi(DOI 名)」、「journal(誌名)」、「volume(巻)」、「pages(ページ番号)」の抽出率が 90%以上であり、これらの項目については出典テンプレートから取得可能である可能性が高い。「year(年)」、「issue(号)」、「author(著者)」の抽出率は、前述の 5 項目に比べると下がるものの、50%以上の割合で記述されていることが分かる。一方で、「publisher」は 1,827 件 (7.1%) であり、抽出率の低い項目である。

issue については、当該の学術情報に号が存在しない可能性がある。author については著者情報自体が書かれていないか、または「first(名)」や「last(姓)」など、author 以外を用いて記述されているために抽出率が低いと考えられる。publisher については、出版者の情報が記述されていないケースが多いことが考えられる。

以上から、DOI リンクを用いた学術情報の参照については、出典テンプレートの情報から、タイトル、誌名、巻、ページ番号は取得可能である割合が高い。その一方で、抽出率が比較的低い、年、号、著者、出版者に関しては、パラメーターのマッピングを検討したうえで、もし、依然として抽出率が低い場合は、Wikipedia のテンプレート情報に依存した方法論の限界であると考えられる。

表 9 の結果から、CiNii の論文詳細ページの参照記述に関して、抽出率の高い書誌要素は「url」と「title(タイトル)」である。url の抽出率は 100%である点については、CiNii の論文詳細ページの抽出処理がこの値の取得のみによって行われているためであり、たとえば「url」ではなく「naid」のパラメーターを用いて記述されている場合は抽出が行われていないと考えられる。

CiNii の論文詳細ページと DOI リンクの参照記述における抽出結果の違いとして、表 8 と表 9 を比較すると、「author」と「publisher」に関しては CiNii の論文詳細ページのほうが抽出率が高い。ただし、CiNii の論文詳細ページにおける「author」の抽出率は 78.4%、「publisher」の抽出率は 34.6%であるため、確かに DOI リンクの参照記述における結果と比較すると抽出率が高いものの、当該の学術情報には著者や出版者が存在することが考えられるため、抽出率改善の余地がある。

以上の分析結果および考察から、Wikipedia の出典テンプレートの情報に依存した方法論としての DBpedia 手法では、リソースごとに書誌要素ごとの抽出率にばらつきがあること、特に、DOI リンクや CiNii の論文詳細ページについては出版者の抽出率が低いことが明らかになった。これらの書誌要素については、それぞれのサービス提供元が提供する Web API を活用する

表 7: すべてのリソースの参照記述における各プロパティの抽出率 (上位 10 件、異なりリソース数: 890,386)

順位	プロパティ	件数	抽出率 (%)
1	http://ja.dbpedia.org/property/title	880,943	98.9
2	http://ja.dbpedia.org/property/url	704,296	79.1
3	http://ja.dbpedia.org/property/accessdate	647,285	72.7
4	http://ja.dbpedia.org/property/publisher	598,603	67.2
5	http://ja.dbpedia.org/property/date	435,207	48.9
6	http://ja.dbpedia.org/property/author	190,807	21.4
7	http://ja.dbpedia.org/property/year	153,067	17.2
8	http://ja.dbpedia.org/property/work	145,171	16.3
9	http://ja.dbpedia.org/property/last	89,336	10.0
10	http://ja.dbpedia.org/property/first	86,085	9.7

表 8: DOI リンクの参照記述における各プロパティの抽出率 (上位 10 件、異なりリソース数: 25,779)

順位	プロパティ	件数	抽出率 (%)
1	http://ja.dbpedia.org/property/title	25,680	99.6
2	http://ja.dbpedia.org/property/doi	25,125	97.5
3	http://ja.dbpedia.org/property/journal	24,969	96.9
4	http://ja.dbpedia.org/property/volume	24,769	96.1
5	http://ja.dbpedia.org/property/pages	24,107	93.5
6	http://ja.dbpedia.org/property/year	20,222	78.4
7	http://ja.dbpedia.org/property/issue	19,184	74.4
8	http://ja.dbpedia.org/property/author	16,246	63.0
9	http://ja.dbpedia.org/property/pmid	11,254	43.7
10	http://ja.dbpedia.org/property/url	6,359	24.7

ことで改善可能であると考えられる。

7 おわりに

本研究では、日本語版 Wikipedia における学術情報の参照を LOD 化するための予備的な分析として、どのような学術情報がどれくらい参照されているかの調査を行ったうえで、参照の多い学術情報であることが明らかになった DOI リンクおよび CiNii URI について、DBpedia 手法と提案手法の比較を通じた分析を行った。

分析の結果、まず、提案手法により、2016 年 3 月 5 日時点の日本語版 Wikipedia の標準名前空間において、DOI リンクが 35,266 件、CiNii URI が 20,418 件記述されていることが明らかになった。FQDN 単位での集計結果から、標準名前空間における外部リンクのうち、DOI リンクは 9 番目、CiNii URI は 13 番目に件数の多い項目である。

次に、DBpedia 手法および提案手法による学術情報の参照結果の比較から、DBpedia 手法について、DOI リンクの抽出率が高いこと、CiNii URI の抽出率が低いことが明らかになった。CiNii の抽出率が低い点については、出典テンプレートに「naid」が記述されている場合はその値を用いた URI の生成を行うなどの対策が考えられる。ただし、抽出処理において、DBpedia 手法では一部の記号が含まれる場合に URI を正確に抽出できない場合がある。これらの結果から、DBpedia 手法を日本語版 Wikipedia にそのまま適用することによって、学術情報の参照を LOD 化するという方法は、必ずしも十分なものとは言いがたい。また、URI の抽出処理における問題点は、日本語版 Wikipedia のみに見られる現象ではないと考えられるため、DBpedia citations & references challenge へのフィードバックを行うことが課題である。

最後に、DBpedia 手法での学術情報の書誌要素ごとの抽出率を分析した結果から、日本語版 Wikipedia の出典テンプレートのパラメーターに依存した手法では、リソースごとに書誌要素の抽出率にばらつきがあり、抽出率の低い項目が含まれていることが分かった。この点については、今後、DOI リンクについては各 RA が提供する Web API、CiNii の論文詳細ページについて

表 9: CiNii の論文詳細ページの参照記述における各プロパティの抽出率 (上位 10 件、異なりリソース数: 1,117)

順位	プロパティ	件数	抽出率 (%)
1	http://ja.dbpedia.org/property/url	1,117	100.0
2	http://ja.dbpedia.org/property/title	1,116	99.9
3	http://ja.dbpedia.org/property/journal	980	87.7
4	http://ja.dbpedia.org/property/pages	917	82.1
5	http://ja.dbpedia.org/property/author	876	78.4
6	http://ja.dbpedia.org/property/volume	828	74.1
7	http://ja.dbpedia.org/property/year	766	68.6
8	http://ja.dbpedia.org/property/issue	509	45.6
9	http://ja.dbpedia.org/property/publisher	387	34.6
10	http://ja.dbpedia.org/property/date	292	26.1

は CiNii の Web API を活用することでデータセットを作成し、DBpedia 手法での書誌要素ごとの抽出結果との比較を通じた詳細な検討を行うことが課題である。

参考文献

[1] 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. http://doi.org/10.2964/jsik_2015_011.

[2] 吉川次郎, 佐藤翔, 高久雅生, 逸村裕. “日本語版および英語版 Wikipedia における DOI リンクの重複分析”. 第 14 回情報メディア学会年次大会. 京都, 2015-06-27. 第 14 回情報メディア学会研究大会発表資料. 2015, p.27-30. http://hdl.handle.net/2241/00125076.

[3] 吉川次郎, 高久雅生, 武田英明, 逸村裕. “アクセスログに基づく DOI リンクの参照状況の分析: JaLC DOI を対象に”. 三田図書館・情報学会 2015 年度研究大会. 東京, 2015-11-14. 2015 年度三田図書館・情報学会研究大会発表論文集. 2015, p.17-20. http://www.mslis.jp/am2015yoko/05_kikkawa_rev.pdf.

[4] Schmachtenberg, Max.; Bizer, Christian.; Jentzsch, Anja.; Cyganiak, Richard. “Linking Open Data cloud diagram 2014”. The Linking Open Data cloud diagram. http://lod-cloud.net/versions/2014-08-30/lod-cloud_colored.svg, (参照 2016-08-18).

[5] DBpedia. “DBpedia citations & references challenge”. Blog – DBpedia. http://wiki.dbpedia.org/blog/dbpedia-citations-references-challenge, (参照 2016-08-17).

[6] GitHub. “dbpedia/extraction-framework: The software used to extract structured data from Wikipedia ” Github. https://github.com/dbpedia/extraction-framework, (参照 2016-08-17).

[7] Tzekou, Paraskevi.; Stamou, Sofia.; Kirtsis, Nikos.; Zotos, Nikos. “Quality assessment of Wikipedia external links”. Proceedings of the 7th International Conference on Web Information Systems and Technologies. http://doi.org/10.6084/m9.figshare.1048991, (参照 2016-08-17).

[8] 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) 情報知識学会年次大会. 香川, 2011-05-28/29. 情報知識学会誌. 2011, vol.21, no,2. p.157-162. http://doi.org/10.2964/jsik.21_06, (参照 2016-08-17).

[9] Nielsen, Finn Arup. “Scientific citations in Wikipedia”. First Monday. 2007, vol.12, no.8, p.1-5. http://doi.org/10.5210/fm.v12i8.1997, (参照 2016-08-17).

[10] Lin, Jennifer.; Fenner, Martin. “An analysis of Wikipedia references across PLOS publications”. altmetrics14 workshop at WebSci. Indiana, 2014-06-23. http://doi.org/10.6084/m9.figshare.1048991, (参照 2016-08-17).

[11] 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. 茨城, 2013-05-25. 2013 年日本図書館情報学会春季研究集会発表論文集. 2013, p.27-30. http://researchmap.jp/?action=cv_download_main&upload_id=46852.

[12] “Submissions/Usage of Digital Object Identifiers across Wikimedia projects”. Wikimania 2015 in Mexico City. https://wikimania2015.wikimedia.org/wiki/Wikimania, (参照 2016-08-18).

[13] Wikipedia. “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 2016-08-18).

[14] Taraborelli, Dario.; Mietchen, Daniel. “Wikipedia Cite-o-Meter: Find citations by publisher in Wikipedia”. Tool Labs. http://tools.wmflabs.org/cite-o-meter/, (参照 2016-08-18).

[15] CrossRef labs. “DOI Chronograph”. CrossRef labs. http://chronograph.labs.crossref.org/, (参照 2016-08-18).

[16] Bilder, Geoffrey. “Geoffrey Bilder: Strategic Initiatives Update”. SlideShare. 2015-11-23. http://www.slideshare.net/CrossRef/geoffrey-bilder-crossref15, (参照 2016-08-18).

[17] The International DOI Foundation. “Key Facts on Digital Object Identifier System”. Digital Object Identifier System. https://www.doi.org/factsheets/DOIKeyFacts.html, (参照 2016-08-17).