SPARQL検索可能なWikipedia構想

An Idea of SPARQLable Wikipedia

小出 誠二 [1*]
Seiji Koide[1]

[1] オントロノミー合同会社
[1] Ontolonomy, LLC

Abstract: Wikipedia is a sort of online encyclopedia on the Internet. The Wikimedia Foundation, Inc, which is an American non-profit and charitable organization, manages and runs every language website of wikipedias. Although they have become huge and popular with people, they also have some shortages for machines in the viewpoint of Semantic Webs and LODs. Aiming to propel the reuse of wikipedias for Semantic Webs and LODs, this paper discusses the pros and cons of Wikipedia, the current status of research and development from the Semantic Web views, and forecasts the future in the direction of Wikipedia for humans to Wikipedia for machines.

1 はじめに

Wikipedia は非営利団体であるウィキメディア財団によって運営・保守され，世界中のウィキペディアンと呼ばれるボランティアによって支えられている，インターネット上のオンライン百科事典である．そのコンテンツは CC-BY-SA 3.0 によって，その規約に従う限り，無料で誰でも何のためにも再利用可能である．その規模の大きさと内容の信頼性において，今日では世界で唯一の存在と言ってよいが，セマンティックウェブ研究者や人工知能研究者から見れば，人間可読ではあるが機械可読としては中途半端であり，その利用は限られている．そのため，これまでも Wikipedia を言語研究のためのコーパスやセマンティックウェブ研究の資源として利用しようとする研究が数多く行われてきた．

Wikipedia から派生した DBpedia[1, 2] は，今日 Linked Open Data (LOD) のハブとしての役割を果たすようになったが，その内容は主には Wikipedia インフォボックスのものであり，text 本文中のページ内容は反映されていない．玉川らによる日本語 Wikipedia オントロジー [3, 4] は text 本文中の内容からも情報を抽出し，オントロジーを構築しようとしたものであるが，残念ながら 2013 年からは更新が途絶えている．

Wikipedia からの情報を多く含む Wikidata を使えば，SPARQL 風な検索により機械利用ができるが，Wikipedia 以外のデータも多く含まれ，しかもそれらが雑多に未整理のまま混在し，特にクラス階層において初心者がその内容を全面的に信用して使うというわけにはいかない．

最近の人工知能ブームを反映して，理化学研究所の革新知能統合研究センターにおいて Wikipedia 構造化プロジェクト [5] が発足し，Wikipedia を対象に機械学習技術を用いた情報の抽出と構造化を行おうとしているが，現時点でセマンティックウェブ研究にどうつながるかは明らかでない．

ここで望まれるのは，Wikipedia 由来であってもよいが，セマンティックウェブ研究者が知識獲得のための信頼できるプラットフォームとして容易に利用可能な，また研究成果の社会的還元の場としての，大規模かつ永続的な知識源である．欧州におけるセマンティック・メディアウィキの活動は現在も進められているが[1]，かつて日本独自のセマンティックウェブ用 Wiki が報告 [6] されたものの，その後日本におけるセマンティックメディアウィキの開発は途絶えている．当時は SPARQL などの検索技術は未熟であって，ハード・ソフトにおいて当時とは比較にならないパワーを利用可能な今日において，SPARQL 利用などを前提としたセマンティックウェブとして利用可能な Wikipedia の実現可能性を探ることは意味があると考える．

2 関連研究

これまで，Wikipedia 関連で多くの研究が行われてきた．本報ではそれらを 1) オントロジー指向，2) コーパス指向，3) 社会実装，の三つの軸で議論する．

*連絡先：オントロノミー合同会社
E-mail: info@ontolonomy.co.jp

[1] https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki

2.1 Wikipedia からオントロジーへ

2.1.1 Wikipedia カテゴリ

Wikipedia のページは様々な観点からカテゴリ化されている．ここでカテゴリとは Wikipedia ページの最後に記載されている情報のことである．たとえば,「徳川家康」(https://ja.wikipedia.org/徳川家康) はカテゴリ「徳川氏」にも「三英傑」にも,「1543 年生」や「1616 年没」にも所属している．Wikipedia カテゴリには上下関係があるが，このカテゴリ上位下位関係はセマンティックウェブで言う包摂概念とは異なることはよく知られており，これをセマンティックウェブの観点で整理する研究が進められている [7].

小林ら [8] は日本語語彙体系における一般名詞意味体系を参考に，日本語 Wikipedia の記事とそのカテゴリの関係からオントロジー構築する手法を示したが，語彙体系における一般語が指す意味の曖昧性に起因する問題を指摘している．

柴木ら [9, 10] は同じく日本語語彙体系のカテゴリに Wikpedia カテゴリ階層を接続する手法で，SVM 分類器を用いてカテゴリ間の is-a 関係を構築したが，必ずしも RDF/OWL 意味論に従わない日本語語彙体系のカテゴリを用いた結果には，包摂関係において再検討の余地がある．

かてて加えて，著作権で保護される日本語語彙体系を用いる場合には，Wikipedia のクリエイティブ・コモンズとしての成果の利用について疑義が残る．WordNet を参考にすれば synset の曖昧性や成果の利用の問題は生じないが，日本語 WordNet においては日本語としての利用には再検討が必要である．森田ら [11] は日本語 Wikipedia オントロジーと日本語 WordNet の統合を行ったが，残念ながらその成果は現状の Wikipedia については適応されていないようである．

2.1.2 Wikipedia インフォボックス

DBpedia は Wikipedia のインフォボックスの情報の内容を RDF にしたものである．ここでインフォボックスとは Wikipedia ページの右側に枠で囲われて記載されている情報のことである．これは Infobox テンプレートや基礎情報テンプレートなどのテンプレートに従った記載によるものであるが，1) 複雑な構造のインフォボックスでは，そのすべてを正確に DBpedia にマッピングできない，2) 日本独自の属性を正しく RDF プロパティにマッピングすることが困難，3) すべてのページにインフォボックスが添付されているわけではない，などの問題があり，特に日本の歴史的，文化的情報に関しては情報抽出に不完全さが残る．一例として「徳川家康」の基礎情報の一部を以下に示す．ここでは,「別名」以外に,「改名」,「神号」,「戒名」などがあることに注目されたい．これらの情報を意味的に知識表現するにはそれなりのオントロジーが必要である．

ソースコード 1: 徳川家康基礎情報

```
{{基礎情報 武士
| 氏名 = 徳川家康／松平元康
| 画像 = Tokugawa Ieyasu2.JPG
| 画像サイズ = 270px
| 画像説明 = 徳川家康像(&#91;&#91;狩野探幽&#93;&#93;画、&#91;&#91;大坂城|大阪城&#93;&#93;天守閣蔵)
| 時代 = &#91;&#91;戦国時代 (日本)|戦国時代&#93;&#93; - &#91;&#91;江戸時代&#93;&#93;前期
| 生誕 = &#91;&#91;天文 (元号)|天文&#93;&#93;11年 &#91;&#91;12月 26日 (旧暦)|12月 26日&#93;&#93;(ユリウス暦 &#91;&#91;1543年&#93;&#93;&#91;&#91;1月 31日&#93;&#93;)
| 死没 = &#91;&#91;元和 (日本)|元和&#93;&#93;2年 &#91;&#91;4月 17日 (旧暦)|4月 17日&#93;&#93;(グレゴリウス暦 &#91;&#91;1616年&#93;&#93;&#91;&#91;6月 1日&#93;&#93;)
| 改名 = 松平竹千代(幼名)→ 元信(初名) → 元康 → 家康 → 徳川家康
| 別名 = 輩行名:次郎三郎&amp;lt;br/&amp;gt;尊称: &#91;&#91;大御所 (江戸時代)|大御所&#93;&#93; (将軍引退後)、&#91;&#91;神君&#93;&#93;(死後) &amp;lt;ref name=&amp;quot;kokusitaijien&amp;quot;&amp;gt;&#91;&#91;尾藤正英&#93;&#93;「徳川家康」『&#91;&#91;国史大辞典&#93;&#93;』&#91;&#91;吉川弘文館&#93;&#93;&amp;lt;/ref&amp;gt;
| 神号 = 東照大権現
| 戒名 = 東照大権現安国院殿徳蓮社崇譽(誉)道和大居士
...
}}
```

一方，オントロジー的観点から見た場合，1) Wikipedia ページタイトルは曖昧性回避が実施されている，2) インフォボックスには一つのテンプレート名あるいはテンプレート種類名が付けられている場合がほとんどであり，これはページタイトルを個物としたときのクラスに相当するとしてよい，などの利点もある．したがって，インフォボックス名とページタイトルをクラス-インスタンス関係と捉え，(ページタイトル-インフォボックス属性名-その値) の関係を (サブジェクト-プレディケイト-オブジェクト) と捉えて RDF 化し，その後，ボトムアップ的に上位クラスの構造化を探索的に行うことが有効と考えられる．

なお，上松ら [12] は DBpedia のエンティティ分類の体系化を目的に，関根の拡張固有表現を用いて DBpedia エンティティを再分類してみた．その結果，関根の拡張固有表現と DBpedia オントロジーに齟齬がありマッチングできない例が多くあることを指摘している．

2.1.3 Wikipedia コンテンツからのオントロジー

Wikipedia のダンプファイルは XML で得られる．ページ ID やコントリビュータなどのページメタ情報は XML から容易に得られるが，表示ページ内容は text

とタグ付けされた中にあり，これらの内容はそれなりに目次化，項目化，箇条書きなどがほどこされているが，それらからどのようにRDFとして情報抽出するかは明らかでない．さらにその内部の平文となると，機械的に情報を抽出するには自然言語処理の技術が必要である．

2.1.4 Wikipedia 知識領域

Wikipedia のページ構成にオントロジー的なガイドラインはなく，ページの中身の構成はもっぱらウィキペディアンのセンスに任されている．オントロジー的には異なる知識領域の問題が一つのページに書かれている場合もあり，Infobox や基礎情報がある場合にはよいが，ない場合にはどのようにページタイトルをクラスづけするかが問題となる．また時には Wikipedia コンテンツを再編成して，知識領域ごとに異なるオントロジーに分割するのがよいと思われる場合もある．その一例として「ゴルディロックスの原理」[2]を挙げておく．

2.1.5 Wikipedia と Wikidata の RDF 化

Wikidata の内容の多くは Wikipedia から来ている．Wikidata を用いれば SPARQL 検索可能という利点はあるが，1) 検索結果の表示が人間向きではない，2)RDF 意味論に従わない，3)Wikipedia に含まれるすべての情報があるわけではない，などの問題がある．RDF 意味論[3]においては，個物と概念が厳密に区別される．rdfs:Resource はいかなる個物もその instance とし，すべての概念の最上位となるものであり，rdfs:Class はすべての概念のクラスとなるメタクラスである．また，クラス-インスタンス関係はプロパティrdf:type により，概念間の上下 (包摂) 関係は rdfs:subClassOf により記述される．Wikipedia は RDF ではないことはよく了解されているが，実は Wikidata も RDF 意味論に従わないという点で RDF ではない．筆者らは RDF/OWL 意味論を高階に拡張し，メタモデリングに関する定式化を行ったが [13, 14]，結論として，メタクラスを含むオントロジーは高階のオーダにより層状化されなければならないとした．Wikidata には Cyc に類似のメタクラス階層が含まれ，このままでは RDF/OWL 推論は不可能である．Erxleben ら [15] は Wikidata の内容の RDF exporting について報告しているし，Brasileiro ら [16] はメタモデリングの観点から Wikidata に含まれるタキソノミー（具体的には “instance of” と “subclass of”）の問題点を指摘した．

[2] https://ja.wikipedia.org/wiki/ゴルディロックスの原理
[3] http://www-kasm.nii.ac.jp/~koide/RDFSemantics-J.htm

2.2 Wikipedia から汎用大規模コーパスへ

Wikipedia ページ内容を自然言語による良質のテキスト資源と見てこれをコーパスとしてアノテーションしようとする試みが，これまで英語ではされてきた.[4,5] しかし，日本語 Wikipedia については本格的な成果は見当たらない．ここで日本語 Wikipedia コーパスとは，日本語 Wikipedia に用いられる語について形態素解析用辞書への登録，テキストへの品詞情報付与アノテーション，およびテキストへの MediaWiki 用マークアップの構文情報アノテーションのことを言う．

Wikipedia 由来の日本語コーパスの位置づけをあらためて以下のように強調する．1) これまでいくつかの日本語コーパスが開発されてきたが有料無料を問わずいずれも著作権が主張され，再利用可能性に言及していないものもある．クリエイティブ・コモンズの観点から問題がある．2) コーパスといえどもその利用目的は様々であり，ゴールオリエンテッドなコーパス構成が望ましい．

我々の考える目的は，将来セマンティックウェブに資するオントロジーに発展することを前提とするものであり，そのような発展を容易とするものであってほしい．たとえば，形態素解析の特徴として，長単語を特徴とするもの，短単語を特徴とするものなどがあるが，ここで考えるのは当然のことながらクラス情報や曖昧性回避の注釈を除いたページタイトルを登録単語とするものである．Wikipedia ページ本文中ではページタイトルには「[[タイトル名]]」といった記述がなされ，ページ表示上では自動的に他のページへのリンクが貼られる．すでにあるこのような Wikipedia の構造は容易にセマンティックウェブのアノテーションに変換することができる．

2.3 社会実装

これまで述べてきたような Wikipedia 由来の大規模オントロジーを実現にするには，Wikipedia と同様にセマンティックウェブなウィキペディアンの協力を可能にすることが望ましい．そのためのしかけと道筋について，衆智を集めて議論したい．

3 むすび

人が読むことを前提に作られてきた Wikipedia は，その規模と内容の確からしさから今日では十分に利用されうるものになっている．しかし，セマンティックウェブや LOD の観点からこれを利用しようとすると，

[4] https://corpus.byu.edu/wiki/
[5] https://www.sketchengine.eu/english-wikipedia-corpus/

それは容易ではない．本報告では，Wikipedia の機械利用について現在の諸問題と研究状況について述べた．自然言語処理技術の発展を待てば，いずれエージェントが Wikipedia コンテンツを理解するようになるかもしれないが，現在そのような未来を実現可能であると楽観的に描くことはできない．一方，SPARQL 検索の有用性は日々実感するところであり，詳細な情報がオントロジーとされなくてもテキスト情報検索技術レベルでの SPARQL 検索が可能になるだけで，最初は一部からオントロジー化とアノテーション付与を初めて，自己進化的な発展が可能になるかもしれない．それを可能とするセマンティックウェブ研究者向けの研究開発用 Wikipedia プラットフォーム構築を目指し，息長い活動を進めるつもりである．

謝辞

本報の内容は革新知能統合センターの森羅プロジェクトと，2018 年 7 月に行われた LOD チャレンジ 2018 ミートアップでのアンカファレンスにおける議論に触発されたものである．此処に記して関係者に感謝の意を表する．

参考文献

[1] Bizer, Christian, et al.: DBpedia - A cristallization point for the Web of Data, *Web Semantics*, Vol. 7, pp. 154–165 (2009)

[2] 加藤 文彦: DBpedia の現在, 情報管理, 60-5, pp.307–315 (2017)

[3] 玉川 奨, 桜井 慎弥, 手島 拓也, 森田 武史, 和泉 憲明, 山口 高平: 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, 25-5, pp.623–636 (2010)

[4] 玉川 奨, 香川 宏介, 森田 武史, 山口 高平: 日本語 Wikipedia オントロジーの構築と利用, 人工知能学会セマンティック Web とオントロジー研究会, SIG-SWOA1203-01 (2013)

[5] 関根 聡, 小林 暁雄, 安藤 まや, 乾 健太郎: 拡張固有表現に基づく Wikipedia 項目の分類と構造化, 人工知能学会セマンティック Web とオントロジー研究会, SIG-SWO-043-05 (2017)

[6] 武田 英明, ムリアディ・ヘンドリー: Semantic MediaWiki の構築に向けて, 人工知能学会セマンティック Web とオントロジー研究会, SIG-SWO-A404-06 (2005)

[7] 中川 嵩教, 吉岡 真治: 知識工学者のための日本語 Wikipedia のカテゴリ階層構造の再整理人工知能学会第 32 回全国大会, 2F4-02, (2018)

[8] 小林 暁雄, 増山 繁, 関根 聡: 日本語語彙体系と日本語ウィキペディアにおける知識の自動結合による汎用オントロジー構築手法, 情報処理学会研究報告, 2008-NL-187 (2), pp.7–14 (2008)

[9] 柴木 優美, 永田 昌明, 山本 和英: Wikipedia からの大規模な人オントロジー構築, 情報処理学会報告, Vol..2010-NL-198, No.3, (2010-9)

[10] 柴木 優美, 永田 昌明, 山本 和英: Wikipedia からの大規模な汎用オントロジー構築, 言語処理学会第 17 回年次大会発表論文集, pp.908–911 (2011)

[11] 森田 武史, 玉川 奨, 山口 高平: 日本語 Wikipedia オントロジーと日本語 WordNet の統合, 第 26 回人工知能学会全国大会, 1—2-R-4-6 (2012)

[12] 上松 大輝, 趙 麗花, Natthawut Kertkeidkachorn, 市瀬 龍太郎: オントロジーマッチングを用いた知識グラフの構築, 人工知能学会研究会資料, SIG-SWO-044-04 (2018-3)

[13] 小出 誠二, 武田 英明: 表示意味論にもとづく RDF/OWL 意味論の形式化, 第 29 回セマンティックウェブとオントロジー研究会, SIG-SWO-A1203-06, (2013)

[14] Inquiry into RDF and OWL Semantics, Joint Int.l Semantic Technology Conf. (JIST2016), Semantic Technology, pp.15–31, Springer (2016)

[15] Erxleben, Fredo and Günther, Michael and Krötzsch, Markus and Mendez, Julian and Vrandeǎić, Denny: Introducing Wikidata to the Linked Data Web, Proceedings of the 13th International Semantic Web Conference - Part I (ISWC '14), pp.50–65 Springer (2014)

[16] Brasileiro, Freddy and Almeida, João Paulo A. and Carvalho, Victorio A. and Guizzardi, Giancarlo: Applying a Multi-Level Modeling Theory to Assess Taxonomic Hierarchies in Wikidata, Proceedings of the 25th International Conference Companion on World Wide Web (WWW '16 Companion), pp.975–980, (2016)