1N4-OS-10b-3

日本語 Wikipedia オントロジーの Linked Open Data への取り組み

An Approach for Linked Open Data of Japanese Wikipedia Ontology

玉川 奨[*1] Susumu Tamagawa　香川 宏介[*1] Kosuke Kagawa　森田 武史[*2] Takeshi Morita　山口 高平[*1] Takahira Yamaguchi

[*1]慶應義塾大学 Keio University　[*2]青山学院大学 Aoyama Gakuin University

This paper presents how to use Japanese Wikipedia Ontology with more property semantics compared with DBpedia, including relationships among other Linked Open Data resources: DBpedia Japanese and LODAC and so on. We also discuss three new properties: jwo:hyper, jwo:nearly, jwo:verb. Japanese Wikipedia Ontology could be popular as a hub on Japanese Linked Open Data.

1. はじめに

大規模なオントロジーは情報検索やデータ統合において有用である．しかしながら，オントロジーの手動構築には，膨大な時間がかかり，保守や更新が困難という問題がある．そこで，近年，オントロジーの自動構築に関する研究は盛んに行われており，その情報資源として，Web 上の百科事典である Wikipedia を利用した研究は多い．Wikipedia は語彙網羅性，即時更新性に優れており，半構造情報資源であることからフリーテキストと比べてオントロジーとのギャップが小さいため，非常に優れた情報資源であるためである．一方で，Linked Open Data(LOD) が国内でも普及し始めている．Linked Open Data チャレンジ Japan[*1]のような普及活動の効果もあり，DBpedia Japanese，LODAC[武田 11]，saveMLAK など，国内でも様々な領域で LOD としてデータを公開・共有する取り組みがなされている．

我々はこれまでにも，日本語 Wikipedia における様々なリソース (カテゴリツリー，一覧記事，リダイレクトリンク，Infobox 等) から，概念および概念間の関係 (is-a 関係，クラスーインスタンス関係，プロパティ定義域，プロパティ値域，プロパティ関係，同義語，インスタンス間関係) を抽出し，高精度かつ大規模な汎用オントロジー (以下，日本語 Wikipedia オントロジー) を学習する手法を提案してきた [玉川 10, 玉川 11]．

本稿では，日本語 Wikipedia オントロジーの Linked Open Data 化へ向けた取り組みについて述べる．

2. 関連研究

DBpedia[Auer 07] は，Wikipedia の半構造情報を RDF に変換することによって，大規模なデータベースを構築している．リソースとしては主に，英語 Wikipedia の Infobox や外部リンク，所属カテゴリといった半構造情報を利用している．LOD のハブとして広く様々な領域の LOD とリンクされている．本家 DBpedia が英語版 Wikipedia を対象にしているのに対し，日本語版 Wikipedia を対象とし，独自でマッピング作業を行なっている DBpedia Japanese[*2]も存在している．

YAGO2[Johannes 10] は YAGO の知識ベースの拡張として，これまでの WordNet に Wikipedia のカテゴリを付加してオントロジーの拡張を行うだけでなく，Wikipedia と GeoNames から時空間的情報を抽出する事で，さらなるオントロジーの拡張を目指している．これら時空間的情報は wasBornOnDate や isLocatedIn といった関係を定義し，インスタンスとつないでおり，非階層関係となっている．非階層関係に着目し，時空間も含めた高度なオントロジーを構築しているが，これらの関係は手動で定義されており，プロパティの定義域や値域についても手動で定義されている．

連絡先: 玉川 奨，山口高平，慶應義塾大学理工学研究科
{s_tamagawa,yamaguti}@ae.keio.ac.jp

3. 日本語 Wikipedia オントロジーの構築

3.1 日本語 Wikipedia オントロジー

図 1 は日本語 Wikipedia オントロジーの概略図である．日本語 Wikipedia オントロジーは以下の関係とタイプから構築される．() 内は，抽出した関係に対応する，OWL[*3]，RDFS[*4]，RDF[*5]，JWO[*6]で定義した語彙を示す．本稿では LOD 化に際し，新たに定義した 3 つの語彙と他の LOD との関連付け方法について述べる．

1. is-a 関係 (rdfs:subClassOf)
2. クラス-インスタンス関係 (rdf:type)
3. プロパティ名とトリプル (以下のプロパティタイプを含む)
   (a) オブジェクトプロパティ (owl:ObjectProperty)
   (b) データタイププロパティ (owl:DatatypeProperty)
   (c) 対称関係プロパティ (owl:SymmetricProperty)
   (d) 推移関係プロパティ (owl:TransitiveProperty)
   (e) 関数関係プロパティ (owl:FunctionalProperty)
   (f) 逆関数関係プロパティ (owl:InverseFunctional Property)
4. プロパティ定義域 (rdfs:domain)
5. プロパティ値域 (rdfs:range)
6. プロパティ上位下位関係 (rdfs:subPropertyOf)
7. 上位下位関係 (jwo:hyper)
8. 関連語・同義語 (jwo:nearly)
9. 動詞とプロパティの関係 (jwo:verb)

*1 Linked Open Data チャレンジ Japan 2012: http://lod.sfc.keio.ac.jp/challenge2012/
*2 DBpedia Japanese: http://ja.dbpedia.org/
*3 OWL: http://www.w3.org/TR/owl-ref/
*4 RDFS: http://www.w3.org/TR/rdf-schema/
*5 RDF: http://www.w3.org/TR/rdf-syntax-grammar/
*6 JWO: http://www.wikipediaontology.org/vocabulary

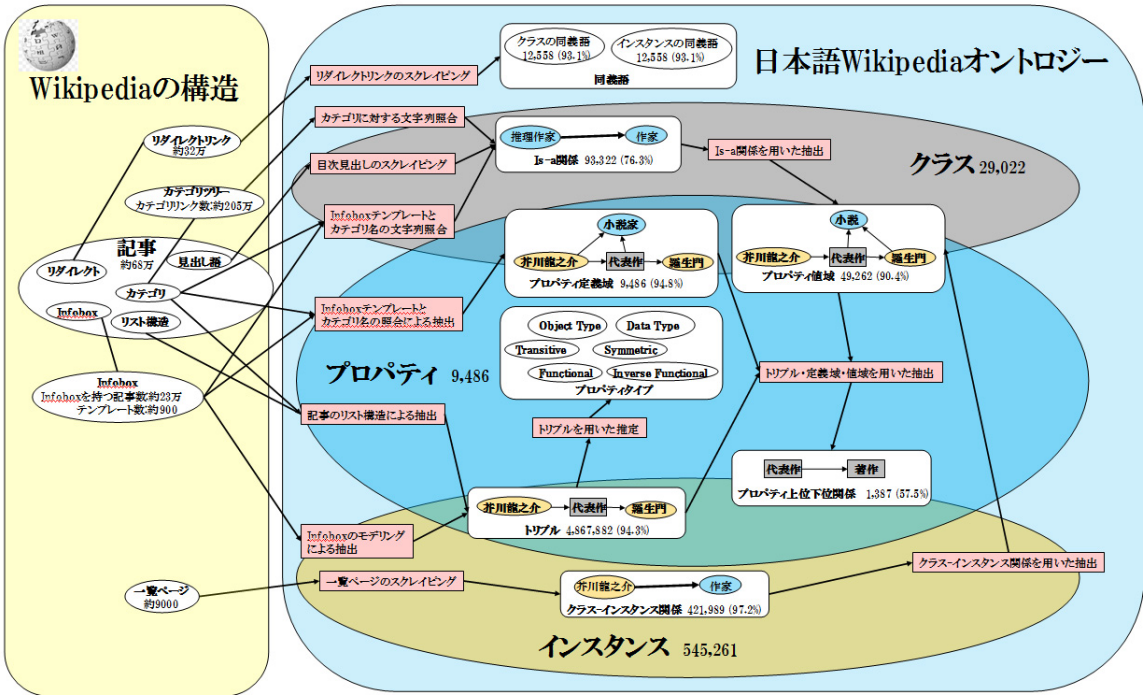

図 1: 日本語 Wikipedia オントロジーの概略図

福澤諭吉

福澤 諭吉(ふくざわ ゆきち、天保5年12月12日(1835年1月10日) – 明治34年(1901年)2月3日)は、日本の武士(中津藩士のち旗本)、蘭学者、著述家、啓蒙思想家、教育者。慶應義塾の創設者であり、専修学校(後の専修大学)、商法講習所(後の一橋大学)、伝染病研究所の創設にも尽力した。他に東京学士会院(現在の日本学士院)初代会長を務めた。そうした業績を元に明治六大教育家として列される。

図 2: 福澤諭吉記事のアブストラクト

3.2 語彙の定義と抽出

3.2.1 上位下位関係の定義 (jwo:hyper)

日本語 Wikipedia オントロジーでは，クラス及びインスタンスを明確に定義していたため，上位下位関係を is-a 関係とクラスーインスタンス関係に分類していた．しかしながら，上位のクラスを持たない記事も多く存在していたため，新たに記事のアブストラクトから上位下位関係を抽出し，ゆるい上位下位関係として，jwo:hyper 語彙により関係を定義した．実際の抽出手順は次のとおりである．

1. Wikipedia 記事の最初の段落をアブストラクトとして抽出
2. いくつかのパターンから記事名を下位語とする上位下位関係を抽出
3. jwo:hyper を語彙として関係を定義

図 2 は福澤諭吉の記事のアブストラクトである．多くの Wikipedia の記事には図のように「記事名 (よみ、生年 - 没年)」は、上位語 1、上位語 2 … 」という記述が見られる．このようなパターンから記事名を下位語として上位下位の関係を構築する．

結果として,「福澤諭吉」記事から「著述家」「蘭学者」,「トヨタ自動車」記事から「自動車メーカー」,「吾輩は猫である」記事から「長編小説」などを上位語として抽出した．

3.2.2 関連語・同義語の定義 (jwo:nearly)

日本語 Wikipedia オントロジーの同義語は Wikipedia のリダイレクトリンクを用いて構築している．これまで同義語として，skos:altLabel を用いて定義していたが，誤りも多く存在しているため，よりゆるいリソース間をつなぐ語彙として jwo:nearly を用いて関係を定義する．また，infobox から直接抽出した infobox プロパティと日本語 Wikipedia オントロジー独自のプロパティの関係も jwo:nearly 語彙により定義する．

結果として,「福澤諭吉」と「福沢諭吉」,「スティーヴジョブス」と「スティーブジョブズ」,「国籍」プロパティと「nationality」プロパティなどを関連語・同意語の関係として抽出した．

3.2.3 動詞とプロパティの関係定義 (jwo:verb)

日本語 Wikipedia オントロジーのプロパティトリプルを用いて，Wikipedia 記事内の文章から同一の目的語が出現する文に注目し，その文中の動詞を抽出する．これにより，プロパティと意味的に近い動詞が抽出できる可能性があり，今後プロパティの表記揺れ問題の対策に利用できる．本関係は jwo:verb 語彙により表記する．例えば，日本語 Wikipedia オントロジーの「周辺情報」プロパティを含むトリプルの目的語は文中で「位置する」「隣接する」といった動詞と共に出現することが多い．こうしたプロパティと動詞を jwo:verb により対応付ける．

結果として，先の「周辺情報」プロパティと「位置する」「隣接する」,「発売元」プロパティと「発売する」「販売する」,「掲載誌」プロパティと「掲載する」などを抽出した．

表 1: プロパティと標準語彙の関連付けの一例

プロパティ名	関連先
人口	gn:population
所在地	gn:locatedIn
近隣，周辺情報	gn:nearby
商品名	gr:name
発売元	gr:Brand
俳優	schema:actor
設立者	schema:founders

3.3 他のリソースとの関連付け

3.3.1 標準語彙との関連付け

日本語 Wikipedia オントロジーのプロパティと schema.org[*7]，GeoNames[*8]，GoodRelations[*9]の語彙を対応付ける．各語彙と意味的に近似している日本語 Wikipedia オントロジーのプロパティを見つけ，owl:sameAs の関係で結ぶ．

例えば，日本語 Wikipedia オントロジーの「国籍」プロパティは schema.org の schema:nationality 語彙と近似である．このような関係を手作業で対応付ける．表 1 に対応付けの一例を示す．

3.3.2 他の LOD との関連付け

日本語 Wikipedia オントロジーのインスタンスと DBpedia Japanese[*10]，LODAC[*11]，青空文庫[*12]，saveMLAK[*13]のリソースの関連付けを行う．日本語 Wikipedia オントロジー内のインスタンスと各 LOD のリソースの文字列照合を行い，完全照合した場合に owl:sameAs によって対応付けを行う．表 2 に関連付けの一例を示す．

4. 日本語 Wikipedia オントロジーの公開

日本語 Wikipedia オントロジーの最新版は 2013 年 2 月時点の Wikipedia ダンプデータ (jawiki-latest-pages-articles.xml)[*14]を利用し，構築したものである．日本語 Wikipedia オントロジーのページ[*15]で閲覧，ダウンロードが可能となっている．図 3 にシステムの概要図，図 4 に最新版の統計情報を示す．LOD として公開するにあたり，RDF ストアとして Virtuoso[*16]を利用しており，SPARQL クエリは Virtuoso を通して結果が返ってくる．各リソースのウェブページとデータはできるだけ，メモリ及びキャッシュに保存するこ

*7 schema.org : http://schema.org/
*8 GeoNames : http://www.geonames.org/
*9 GoodRelations : http://www.heppnetz.de/projects/goodrelations/
*10 DBpedia Japanese : http://ja.dbpedia.org/
*11 LODAC : http://lod.ac/
*12 青空文庫 : http://www.aozora.gr.jp/
*13 saveMLAK : http://savemlak.jp/

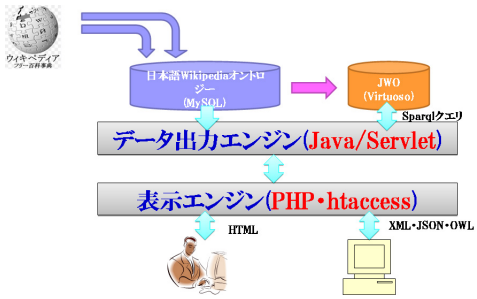

図 3: 日本語 Wikipedia オントロジー LOD のシステム概要図

統計情報	
クラス数	160,248
インスタンス数	1,836,958
プロパティ数	24,382
クラスを持つインスタンス数	701,384
is-a関係数(rdfs:subClassOf)	56,963
タイプの数(rdf:type)	1,009,542
定義域関係数(rdfs:domain)	29,212
値域関係数(rdfs:range)	70,844
プロパティ上位下位関係数(rdfs:subPropertyOf)	302
上位下位関係数(jwo:hyper)	271,834
関連語・同義語(jwo:nearly)	256,517
動詞とプロパティの対応数(jwo:verb)	60,519
プロパティトリプル数	9,820,069
Infoboxトリプル数	2,883,133
外部への参照数(owl:sameAs)	982,907

図 4: 日本語 Wikipedia オントロジー統計情報 (20130216 版)

とで高速に表示するようにしている．

4.1 URI の定義

日本語 Wikipedia オントロジーの URI は表 3 の通りである．

各リソースは「/」以下に日本語もしくは URI エンコードされた日本語を入力することでアクセス可能である．301 リダイレクトにより，ブラウザからのアクセスは「page」へ，アプリケーションからのアクセスは「data」へアクセスする．現在選択できるデータの種類は rdf，owl，rdf/json の 3 種類である．例えば，「http://www.wikipediaontology.org/instance/福澤諭吉」へブラウザからアクセスした場合は「http://www

*14 Wikipedia ダンプデータ: http://download.wikimedia.org/jawiki/
*15 日本語 Wikipedia オントロジー: http://www.wikipediaontology.org
*16 Virtuoso: http://virtuoso.openlinksw.com/

表 2: 他の LOD リソースとの関連付けの一例

日本語 Wikipedia オントロジー URI	関連先 URI
http://www.wikipediaontology.org/instance/福澤諭吉	http://ja.dbpedia.org/resource/福澤諭吉
http://www.wikipediaontology.org/instance/福澤諭吉	http://www.aozora.gr.jp/index_pages/person296.html
http://www.wikipediaontology.org/instance/吾輩は猫である	http://www.aozora.gr.jp/cards/000148/card789.html
http://www.wikipediaontology.org/instance/ギアナウズラ	http://lod.ac/species/ギアナウズラ
http://www.wikipediaontology.org/instance/慶應義塾普通部	http://savemlak.jp/wiki/慶應義塾普通部
http://www.wikipediaontology.org/instance/東京都立大島高等学校	http://savemlak.jp/wiki/東京都立大島高等学校
http://www.wikipediaontology.org/instance/落穂拾い	http://lod.ac/id/497029

表 3: 日本語 Wikipedia オントロジー URI

リソース		URI
インスタンス	URI	http://www.wikipediaontology.org/instance/
	ページ	http://www.wikipediaontology.org/pages/instance/
	データ	http://www.wikipediaontology.org/data/instance/
クラス	URI	http://www.wikipediaontology.org/class/
	ページ	http://www.wikipediaontology.org/pages/class/
	データ	http://www.wikipediaontology.org/data/class/
プロパティ	URI	http://www.wikipediaontology.org/property/
	ページ	http://www.wikipediaontology.org/pages/property/
	データ	http://www.wikipediaontology.org/data/property/
infobox プロパティ	URI	http://www.wikipediaontology.org/infobox/
	ページ	http://www.wikipediaontology.org/pages/infobox/
	データ	http://www.wikipediaontology.org/data/infobox/

図 5: 検索実行結果の一例

```
SELECT ?prop ?value
WHERE
{
  <http://www.wikipediaontology.org/instance/福澤諭吉> ?prop ?value.
}
```

URL
http://www.wikipediaontology.org/query/?q=SELECT+%3Fprop+%3Fvalue%0D%0AWHERE%0D%0A{%0D%0A+++%3Chttp%3A%2F%2Fwww.wikipediaontology.org%2Finstance%2F%E7%A6%8F%E6%BE%A4%E8%AB%AD%E5%90%89%3E+%3Fprop+%3Fvalue.%0D%0A}%0D%0A%0D%0A&type=xml&limit=100

図 6: SPARQL クエリの一例

.wikipediaontology.org/pages/instance/福澤諭吉」へリダイレクトされる．福澤諭吉の URI エンコードである「%E7%A6%8F%E6%BE%A4%E8%AB%AD%E5%90%89」へアクセスした場合も同様である．

4.2 検索実行画面

図 5 が検索実行結果の一例である．入力語に完全一致するリソース，部分一致するリソース，入力語を目的語とするリソースを関連候補として，順に表示している．

4.3 SPARQL エンドポイントの公開

SPARQL エンドポイントは http://www.wikipediaontology.org/query/である．図 6 上部のような SPARQL クエリを投げる場合，図 6 下部のような URL にアクセスすることで，xml 形式でデータを得ることができる．

5. おわりに

本稿では，日本語 Wikipedia オントロジーの Linked Open Data への取り組みについて述べた．日本語 LOD の更なる普及のためには，ハブとなる LOD のリソースの充実は必要不可欠である．DBpedia Japanese のサービスが開始されているが，我々の日本語 Wikipedia オントロジーも DBpedia Japanese と共に補完し合う形で共存できると考えている．

今後も，LOD を中心とした日本語 Wikipedia オントロジーの利用法を検討していく一方，より実用性を考慮し，オントロジーとしての質の向上と規模の拡大を行っていく予定である．なお，日本語 Wikipedia オントロジーは日本語 Wikipedia オントロジー研究ページ*17で公開中であり，検索等が可能である．また，研究ページでは LOD を利用したデモアプリケーション等も公開中であり，興味のある方は参照されたい．

参考文献

[Auer 07] Soren Auer,Christian Bizer,Georgi Kobilarov,Jens Lehmann1,Richard Cyganiak,Zachary Ives: DBpedia: A Nucleus for a Web of Open Data ,Lecture Notes in Computer Science ,Springer Berlin / Heidelberg ,pp.722-735(2007)

[Johannes 10] Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Gerhard Weikum: YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia, Research Report MPI-I-2010-5-007, Max-Planck-Institut für Informatik(2010)

[玉川 10] 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 論文特集「2009 年度全国大会近未来チャレンジ」 Vol.25 No.5 pp.623-636 (2010)

[玉川 11] 玉川 奨, 森田 武史, 山口 高平, ”日本語 Wikipedia からプロパティを備えたオントロジーの構築”, 人工知能学会論文誌 特集論文「近未来チャレンジ」 Vol.26 No.4 pp.504-517 (2011)

[武田 11] 武田 英明, 嘉村 哲郎, 加藤 文彦, 大向 一輝, 高橋 徹, 上田 洋, ”日本における Linked Data の普及にむけて”, 人工知能学会全国大会 (第 25 回) 論文集, No.3E3-OS20-9 (2011)

*17 http://www.wikipediaontology.org/