WordNet 日本語化への LOD アプローチ

An LOD Approach toward WordNet Japanization

小出誠二 [1*]　武田英明 [2]　大向一輝 [2]
Seiji Koide,[1]　Hideaki Takeda,[2]　Ikki Ohmukai[2]

[1] 情報・システム研究機構新領域融合研究センター
[1] Research Organization of Information and Systems
[2] 国立情報学研究所
[2] National Institute of Informatics

Abstract: We had transformed English WordNet into RDF according to the guideline from W3C and provided RDF files and the way of web browsing of WordNet. While Japanized WordNet is recently published, it is not in the LOD style and does not include links to other existing resources of Japanese word collections and dictionaries. Therefore, RDFization and building LOD as a whole of English/Japanese multi-lingual WordNet is an urgent requirement in today's situation. We also have gained several experiences of LOD in other fields, and the experiences are revealing a new question that should be categorized into pragmatics of RDF that teaches how to RDFize resources and data in each particular context and objective. In this report, we discuss how to RDFize English and Japanese WordNet and how to extend RDFized WordNet to LOD with other existing resources of Japanese words.

1　はじめに

筆者らはすでに W3C の指針 [1, 2] に従って英語 WordNet の RDF 化を行って RDF ファイルを公開し [3]，さらに日本語 WordNet を IPADIC や Wikipedia のような他言語資源と一緒に Web ブラウズ[1]する手段も提供した [4]．一方，LOD の動きが活発化するにつれて，Web 資源や非 Web 資源をどのように RDF 化したらよいのかについて，各所で試行錯誤が行われつつあり，その結果，これまでの RDF 統語論と RDF 意味論に加えて，新たに RDF 語用論とでもいうべき分野のあることが見えてきた．本報では，英語/日本語 WordNet の RDF 化の経験を踏まえて，WordNet の構築とその他辞書の RDF 化について LOD の視点から議論する．

第 2 節では英語 WordNet の構造と特徴について簡単に述べ，W3C による RDF 化の指針についても振り返る．第 3 節ではその後の英語/日本語 WordNet の改定とそれへの対応について述べ，その他の関連研究についても述べる．第 4 節ではそれまでの議論を踏まえて日本語資源の LOD 化について提案する．第 5 節では LOD 化において特に日本語について考慮すべき諸課題について議論し，最後に第 6 節でむすびとする．

*連絡先：国立情報学研究所
〒 101-8430 東京都千代田区一ツ橋 2-1-2
E-mail: koide@nii.ac.jp
[1]ことはぶ，http://wordnet.jp/kotohub/

2　英語 WordNet の構造と W3C による RDF 化指針

RDF 化に当たっては対象分野の構造をよく見極めなければならない．英語 WordNet[6] の特徴は，多義語と同義語を考慮して同義語集合である synset を中心に据えて．単語ならびに複合語と synset の関係を収集整理したことである．例えば，単語 “bank” には「機能としての銀行」のほかにも,「建物としての銀行」や,「土手」や「飛行機の傾き」などの意味がある．一方,「機能としての銀行」の同義語には “bank” のほかに “depository financial institution” や “banking company” とか “banking concern” の複合語がある．Sense key というものが WordNet にはあるが，このような語と同義語集合の多対多の関係を RDF 記述するにあたり，W3C の RDF 化指針では語義 (wordsense) という概念を導入して語と同義語集合を関係づけた．例えば「機能としての銀行」は，単語 “bank” とある語義によりつながっているが，単語 “bank” は異なる語義を介して同義語集合「土手」や同義語集合「飛行機の傾き」にもリンクされている．

まず最初にリソースは URI 表現されなければならない．このとき **URI は曖昧さなく他のリソースと区別され解釈が唯一となるように URI 表現されなければならない．**W3C による RDF 化指針では，WordNet におけ

る単語 “bank” は版も考慮して “wn20instances:word-bank” と URI 表現され (ここで wn20 は WordNet バージョン 2.0 を表す)，synset は同義語集合の中から代表名 (primary name) が選ばれて，“wn20instances:synset-depository_financial_institution-noun-1” と URI 表現[2]される．英語は一つの単語が異なる品詞で用いられることがあるため，word には URI 表現に品詞情報は付かないが synset では付加される．

word ノードと synset ノードの多対多の関係を形作る wordsense ノードをグラフ中で唯一に URI 設定するのは単純な話ではない．簡単には順にシステム中に両者の関係を追加するときに追番によって作ることもできるが，もし何らかの体系化された方法を採るとするとそのアルゴリズムが要求される．W3C による RDF 化指針ではある方法が採用されたがそのアルゴリズムまでは詳細には記述されていないため，スクラッチで実装しようとすると実際には頭を悩ませることになる．

RDF グラフ中の二つのノードをリンクするために新たにプロパティ定義する場合にはできるかぎりプロパティの定義域と値域を考慮して定義するのが望ましい．W3C による RDF 化指針では word インスタンスと wordsense インスタンスはプロパティ wn20schema:sense とその逆リンク wn20schema:word でリンクされ，WordSense インスタンスと Synset インスタンスはプロパティ wn20schema:inSynset と逆リンク wn20schema:contains WordSense でリンクされた．

英語 WordNet には品詞と意味のほかにも IS-A 関係を表す hypernym/hyponym や PART-OF 関係を表す holonym/melonym などがあるが，これらの情報についても W3C による RDF 化指針ではスキーマ定義がされ，wnfull.rdfs[3]となっている．なお，hypernym/hyponym 関係は rdfs:subClassOf のサブプロパティとはなっていないため，RDFS と OWL 意味論においてクラスの上下関係にはなっていないことを注意しておく．

3 英語/日本語 WordNet の改定と RDF 化

3.1 英語 WordNet の改定

W3C の RDF 化指針は WordNet2.0 に対するものであったが，その後英語 WordNet は 2.1，3.0 と改定を重ねて現在に至っている．WordNet2.1 では新たに instanceHypernym と instanceHyponym がインスタンス関係として追加され，それに伴って 2.0 の内容の一部も変更された．WordNet3.0 では新たな関係はない．

WordNet の改定に追随するにあたり，以前はスキーマ定義におけるすべてのスキーマについても wn20schema の名前空間から wn21schema，wn30schema へ更新していたが，新たなスキーマ定義を極力さけて流用できるものは流用するという考え方から現在は WordNet2.0 のための wnfull.rdfs をそのまま 2.1 および 3.0 にも流用して，wn21schema:instanceHyponymOf/instanceHypernymOf のみを定義した．一方，WordNet におけるすべての Word と WordSense と Synset のインスタンスは，たとえ内容が更新されていなくても WordNet2.1 や 3.0 での内容という意味で，wn20instances から wn21instances と wn30instances の名前空間に更新した．

3.2 関連研究

W3C の RDF 化指針以前の WordNet の RDF 表現については，先の報告 [7] に纏めてある．W3C の RDF 化指針は Working Draft であるが，その詳細は van Assem[2] に詳しい．de Melo と Weikum[8] は多言語における語や文字の関係を LOD 化することを目的に「ことはぶ」と類似の環境を公開[4]している．van Assem らは W3C の RDF 化指針に沿って実際に WordNet3.0 に対して RDF 化を行った[5]．

3.3 日本語 WordNet

NICT による日本語 WordNet[5] は英語 WordNet をベースにその synset について日本語を付与したものである．したがって日本語 “銀行” や日本語 “バンク” が sense を介して英語の synset につながっており，それを忠実に RDF 化すると，synset-銀行は生じない．また hypernym/hyponym 関係や holonym/melonym 関係なども英語のままである．英語では単語としてある概念が日本語にはない場合もあるが，単語 “bank” の一つの意味である「賭場ないしは賭事筋の仲介業者が持っている資金」にも日本語の “銀行” という語が振られている．

4 英語/日本語 WordNet の LOD 化

NICT による日本語 WordNet は英語 WordNet 由来であるため，英語を母国語とするユーザの場合には問題がないが，日本語を母国語とするユーザにとっては違和感がある．たとえば，単語 “バンク” に動詞があり，synset-chopstick には wordsense-はし-noun がリンクされるが，その日本語 gloss は「食物を食べる東洋の食器

[2]冗長になるのを避けるために本報ではすべての URI を QName で記す．

[3]http://www.w3.org/2006/03/wn/wn20/download/

[4]http://www.lexvo.org/

[5]http://semanticweb.cs.vu.nl/lod/wn30/

類として使われる一対の細長い棒のうちの 1 本」となる．一方，“火箸” は含まれているが “箸” はない．言語資源として日本語母語のための日本語 WordNet が必要であることはあきらかであり，自由に利用できる資源として IPADIC や NAISTjdic に注目する．言語資源の Web ブラウジング環境として「ことはぶ」を提供しているが，ここでは IPADIC 辞書その他の言語資源も利用して単語に加えて「読み」つながりでブラウジング可能になっていても，RDF 化はされていない．IPADIC あるいは NAISTjdic から出発してこれを RDF 化し，英語 WordNet とリンクさせることで，本来の英語/日本語 WordNet の LOD になると期待できる．

先に見たように，英語概念と日本語概念の構造は異なっていることが予想されるが，それを解決する出発点としてまず最初に英語 synset とは別に日本語 synset を設定する．英語 synset のコピーとしての日本語 synset には問題点の多いことは上記のとおりであるが，**不完全ながらも LOD 化し公開**したのち，それを Web2.0 的に改善していくというのが我々の LOD アプローチである．

5 LOD 化における課題，RDF 語用論

5.1 資源の IRI 化

URI は IETF の RFC3986 で規定されており，さらに現在では URI の国際化を目的に IRI が RFC3987 で決められている．IRI では URI における US-ASCII 文字中の unreserved 文字としての US-ASCII 文字コードが UCS(ISO10646 および unicode) まで拡張された．reserved 文字は URI と IRI では変わらないので表現文字において URI は IRI に包含される．RFC3987 では xsd:anyURI は IRI を含む，ただし http プロトコルには IRI は許されない，とされている．

RDF の勧告としてはまだ URI に留まっているが，流れとしては IRI 化にあり，近い将来 IRI が推奨されるであろう．URI に日本語を用いた場合の問題は，URI に用いられる非 ASCII 文字をパーセントエンコーディングしなければならず，人にとって日本語の視認性に欠けることである．日本語資源を LOD 化する場合，IRI や QName に日本語を記述できることが望ましいのは，RFC3986 においても「URI は、しばしば人間が記憶しなければならないが、その構成要素が意味を持つ、あるいは親しみやすものから成る場合は、人間にとってより覚えやすい」と記されているとおりである．

本来 IRI ではホスト部，パス部，フラグメント部のいずれにも日本語が可能であるが，それと http プロトコルとは別の問題である．将来 IRI が普及した場合でも，RFC3987 では http プロトコルにおいては IRI は許されず，IRI presentation elements を URI protocol elements に変換して通信路に乗せたり，受け取った URI protocol elements を IRI presentation elements に戻したりする操作が処理系には要求される．現在でもファイルとしての形態において RDF を IRI 記述にして維持することに何ら問題はないが，上記のように http プロトコルにおいては日本語は許されない．将来の IRI 化を見越して今から LOD 化をする場合，**現状では URI のフラグメントには日本語を許すことにして，パス部は US-ASCII 文字にするというのが一つの妥協点**である[6]．通常の http プロトコルにおいてフラグメント部が現れることはないため，コンテンツネゴシエーションで rdf フォーマットを扱っても問題は生じない．ただし SPARQL においてはフラグメント部まで含めて扱われるため，現状では問題があると言わざるを得ない．SPARQL のクエリフォームに日本語を与えた場合にはそれを自動的にパーセントエンコーディングする処理系が望まれる．

現在でも rdfs:label や rdfs:comment を日本語記述することは問題ないが，xml:lang を用いる場合には注意が必要である．RDF/XML 記述においては xml:lang が現れると通常のプログラムにおける静的スコーピングと同様に扱われる．すなわち，出現以降の内側にネストされた部分と同レベルにおいても出現以降現れる要素が影響を受ける．すなわち RDF/XML 記述には順序がある．ところがこれを順序のない三つ組にした場合，正確には日本語部分は言語タグつき文字列に変換されなければならないが，すべての処理系でそう処理されるかどうかは心もとない．xml:lang がない場合には日本語であってもそのまま言語タグなし文字列となる．

5.2 プロパティの設定

RDF は誰もが何についても発言できる (Anyone can say anything about anything) ことをその根本精神としている．RDF はプロパティオリエンテッドな知識表現言語である．すなわち OWL とは異なって，何よりも先に，オブジェクトなしにプロパティを定義できるし，他人の定義したサブジェクトやオブジェクトを流用して自分が勝手に異なるプロパティでリンクすることさえ，意味論的に可能になっている．実際には自分のあるいは他人の作ったリソース URI が豊富にあったとして，それらをどうリンクするかというところで大いに悩む．RDF(S) および OWL のプロパティに加えて，**FOAF，Dublin Core, SKOS などの既存のプロパティで間に合う場合にはそれらを流用する**が，さ

[6] ホスト部の日本語 JP ドメイン名は http プロトコルにおいて punycode 化されるが，日本語 JP ドメイン名が一般的に普及しているとは言えない．

もないと特にこれまで RDF 化されていない分野においては自分でプロパティを設定せざるを得ない．

他人の定義したプロパティを用いるということはそれらの定義域と値域の定義にしたがうということであり，リンクされたサブジェクトおよびオブジェクトはそれらのクラスのインスタンスとなる．プロパティを自分で定義する場合，その定義域と値域には最新の注意が必要である．というのは RDF においてはインスタンスは複数のクラスに所属できるため，それ自体が禁止されるわけではないが，定義域と値域を指定することで現在あるいは将来に解釈が成立しない (互いに disjoint なクラスが両立した場合) 可能性があるからである．新たなプロパティを定義する場合には，**特に主張がなければあるいは最初は，定義域と値域を定義しない**というのも現実的な方法かもしれない．

5.3 典拠の問題

Web 資源，非 Web 資源を LOD 化する場合にはその情報の由来すなわち典拠 (provenance) が重要となる．これは特に現在，権利関係にからんで美術館・博物館の LOD 化において強い要求があるが，情報の信頼性などから他分野においても同様である．

典拠の記述方法として，大きく二つの手法がある．一つは現在の三つ組構文は変えずに，その構造の枠内で新たな典拠記述の枠組みを定式化しようとするもの，もう一つは三つ組構文を拡張して典拠情報も記述しようとするものである．Named Graph[9, 10] では任意の三つ組集合に名前を付けることができ，その名前をそれあるいは他の三つ組集合の内部で引用することもできる．

典拠問題は LOD 化された情報のバージョン管理の面でも欠かせない問題である．特に Web2.0 的に公開して三つ組の修正更新を許す場合には，バージョン管理が必須の機能になる．W3C Provenance Working Group では最近 Time Line もデータモデルに含む Working Draft[7]を出した．

6 むすび

英語/日本語 WordNet の RDF 化について報告しつつ，LOD 化にあたっての諸問題を述べた．それらは本文中太字に記述して条項とした．現状特に日本語について LOD 化の問題点が多くあり，欧米の研究者には理解されない部分でもあって，もどかしいところであるが，CJK はもちろんアラビア語，キリル語まで含めれば，セマンティックウェブの国際化は必須であり，これらの問題は解決されなければならない．LOD における日本語問題はその突破口であり，今後も解決の一助となるべく努力する．

[7] http://dvcs.w3.org/hg/prov/raw-file/default/model/ProvenanceModel.html

謝辞

本報で述べたことのほとんどは，NII の LODAC プロジェクトにおける議論から生まれたものである．LODAC チームの全員に感謝する．

参考文献

[1] van Assem, M., Gangemi, A., Schreiber, G., (eds.): RDF/OWL Representation of WordNet, W3C Working Draft 19 June 2006, http://www.w3.org/TR/2006/WD-wordnet-rdf-20060619/

[2] van Assem, M., Gangemi, A., Schreiber, G.: Conversion of WordNet to a standard RDF/OWL representation, Proc. LREC (2006)

[3] Koide, S., Morita, T., Yamaguchi, T., Muljadi, H., Takeda, H.: RDF/OWL Representation of WordNet 2.1 and Japanese EDR Electric Dictionary, 5th International Semantic Web Conference (ISWC2006), Poster (2006)

[4] 丹, 大向, 武田: 日本語リポジトリ「ことはぶ」の構築, 第 24 回人工知能学会全国大会, 2B1-2 (2010)

[5] Isahara, H., Bond, F., Uchimoto, K., Utiyama, M., Kanzaki, K.: Development of Japanese WordNet, In LREC-2008, Marrakech (2008)

[6] Fellbaum, C.: WordNet: An Electronic Lexical Database, MIT Press (1998)

[7] 小出ほか: WordNet と EDR の OWL 表現第 13 回セマンティックウェブとオントロジー研究会, SIG-SWO-A601-03 (2006)

[8] de Melo, G., Weikum, G.: Language as a Foundation of the Semantic Web, (ISWC 2008 Posters & Demos) (2008)

[9] Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Named Graphs, Provenance and Trust, Technical Report HPL-2004-57, (2004)

[10] Carroll, J.J., Bizer, C., Hayes, P., Stickler, P.: Semantic Web Publishing using Named Graphs, Notes of Workshop on Trust, Security, and Reputation the Semantic Web, ISWC-2004, (2004)