1G5-OS-19b-7

日本語 Linked Data Cloud の現状

Current Status of Japanese Linked Data Cloud

加藤 文彦 *1 Fumihiro Kato　武田 英明 *2 Hideaki Takeda　小出 誠二 *1 Seiji Koide　大向 一輝 *2 Ikki Ohmukai

*1情報・システム研究機構 Research Organization of Information and Systems

*2国立情報学研究所 National Institute of Informatics

In recent years, the Japanese Linked Data community has grown up and the number of Japanese datasets published as Linked Data has increased. This paper illustrates the Japanese Linked Data Cloud diagram and describes the current status of Linked Data in Japan.

1. はじめに

Linked Data [Berners-Lee 06] についての説明でよく使われる図として，The Linking Open Data cloud diagram [Cyganiak 11] (以下，本家図) がある．本家図は 2007 年 5 月に最初の版が発表されて以来，2011 年 9 月まで更新されており，Linked Data の発展や現状を紹介するための端的な図として広く用いられている．2011 年 9 月の段階では本家図内に 295 のデータセットが描かれている．

本家図における 1 つの課題は，日本のデータ公開者による日本語のデータセットが ndlna と NDL subjects のみであることである．これらはどちらも国立国会図書館が公開しているデータであり，実際は Web NDL Authorities *1 として統合されているため，1 つのデータセットのみが 2011 年 9 月の本家図に存在すると言える．

一方で，DBpedia Japanese [Kato 13] や日本語 Wikipedia オントロジー [玉川 13] のように，ここ数年で日本においても様々なデータセットが Linked Data として公開されるようになってきており，人工知能学会 セマンティック Web とオントロジー研究会や，2011 年から 3 回開催されている Linked Open Data チャレンジ Japan*2 等においても数多く報告されている．そこで本稿では，日本で公開されている日本語のラベルを含んでいる Linked Data についてのリンク関係等を調査することで，日本語 Linked Data Cloud 図 (以下 JLDC 図) としてまとめることを試みた．

2. 日本語 Linked Data Cloud 図

2014 年 3 月 10 日現在での JLDC 図は図 1 の通りである．現在対象となっているデータセットの数は 27 個である．調査方法は手動であり，過去のセマンティック Web とオントロジー研究会や Linked OpenData チャレンジ Japan にて報告されているデータセットを主な調査対象としている．SPARQL エンドポイントがある場合は SPARQL による問い合わせを行い，RDF ファイルが取得可能な場合はダウンロードしてローカルで計測している．

JDLC 図に採用するデータセットの基準は以下の通りである．

- データ公開者が日本にいる人・組織等である
- 日本語ラベルを含んでいる
- 1000 トリプル以上含んでいる
- 本家図か JLDC 図の既存のデータセットとの RDF リンクが 10 以上ある
- 参照解決可能な状態，データダンプ，あるいは SPARQL エンドポイントのいずれかによってデータセットを公開している

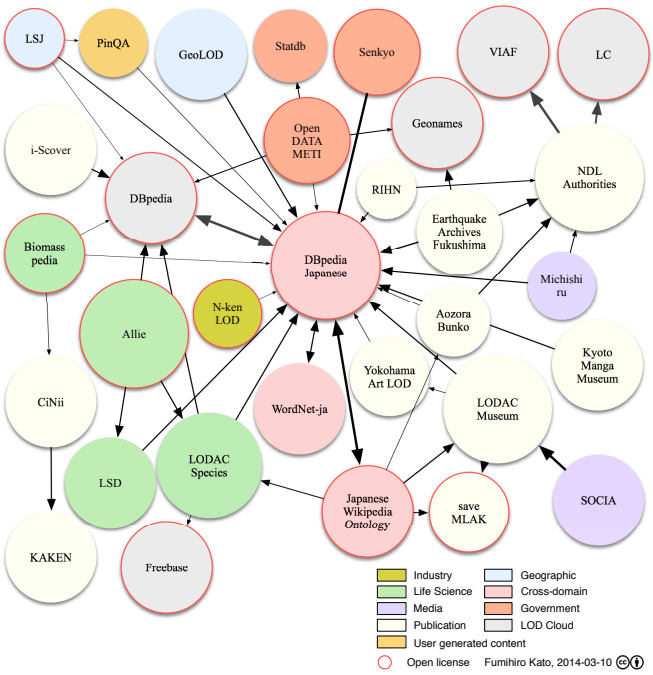

図 1: 日本語 Linked Data Cloud 図

本家図に入っているデータセットについても，本基準に合致するものは JLDC 図のデータセットとして扱うことにした．その理由は，将来 JLDC 図にしかなかったデータセットが本家図に含まれるようになったときにも変わらずに JLDC 図内でも描けるようにするためである．現在該当するのは Web NDL Authorities のみとなる．

連絡先: 加藤 文彦，情報・システム研究機構，東京都千代田区一ツ橋 2-1-2，03-4212-2658，fumi@nii.ac.jp

*1 http://id.ndl.go.jp/auth/ndla/

*2 http://lod.sfc.keio.ac.jp

JLDC 図において，Open Definition [*3] に適合するオープンライセンスを採用しているデータセットは赤丸で明示している．また，データセットの分類は本家図を参考に独自に行った (表 1, 図 2)．唯一，ねじ LOD [Fujii 13] のみは本家図のカテゴリでは分類できないと考えたため，Industry という分類を新設した．

JLDC 図の基準に近いが採用しなかったものとして，RDF リンクの作法が間違っているデータセットが存在した．RDF リンクが間違っているケースとして良くあるのが，Cool URIs [Ayers 08] に関するものである．Cool URIs においてはリソース識別子としての URI とそのリソースに関する文書の URI を明示的に分けることが求められている．しかし，そのように実装されたシステムに対するリンクを生成する際に，文書，特に HTML 文書の URI を使用してしまう例が見受けられた．これは恐らくブラウザでリソースにアクセスした際に，アドレスバーにある URI をそのままリンクに使用したことが考えられる．

3. 本家図の採用基準

JLDC 図におけるデータセットの採用基準は，本家図の採用基準とは異なる．そうした理由は，本家図の基準をそのまま適用すると基準に満たないデータセットが多かったからである．本家図の採用基準をそのまま用いるよりも，まず RDF としてのリンク関係があるデータセットを拾えるようにすることで，個々に何が不足しているのかを示すことが重要だと考えた．

本家図の採用基準 [Cyganiak 11] は Linked Data の 4 原則 [Berners-Lee 06] を解釈したものであり，以下の通りである．

1. 解決可能な http://(または https://) URIs でなければならない
2. content-negotiation か何かで良く使われる RDF 形式 (RDFa, RDF/XML, Turtle, N-Triples) のいずれかで RDF データを解決できなければならない
3. 1000 トリプル以上含んでいる
4. 本家図上の既存データセットとの RDF リンクが 50 以上必要である．
5. RDF クローリングまたは RDF ダンプ，あるいは SPARQL エンドポイントによってデータセット全体にアクセスできる．
6. 認証なしかつ無料でアクセスできる．

最後の項目は基準として記述されているのではなく，オープンについての説明で前提として記述されていることであるが，JLDC 図で本項目に満たないものがあるため基準として明示した．本家図では，オープンは Open Definition に適合するオープンライセンスを採用しているという意味ではなく，認証なしかつ無料でアクセス可能であることとしている．これは，現実にはライセンスを明示的しているデータセットというのが少なかったという事情を考慮したものである．

本家図の基準に合致するデータセットだけを残して図を描くと図 3 のようになる．採用データセット数は 13 個と，JLDC 図の半分以下である．

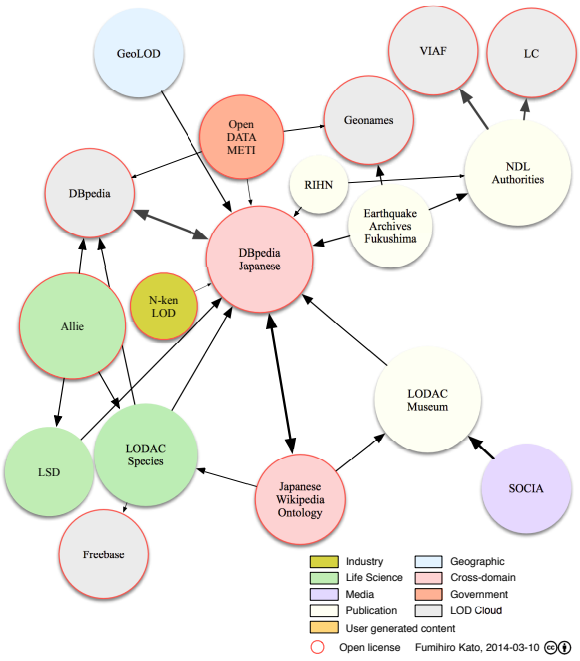

図 3: 本家図基準

基準	データセット数
1 解決可能な http URIs	8
2 RDF データの解決	9
3 1000 トリプル以上	0
4 50 以上の RDF リンク	4
5 データセット全体へのアクセス	2
6 認証なしかつ無料のアクセス	1

表 2: 基準を満たさないデータセット数

基準を満たしていない 14 個のデータセットについて，満たしていない基準毎に表 2 としてまとめた．ここでは複数の理由を許している．

該当数が多い理由が 1 と 2 であるが，これらは Linked Data の 4 原則における 2 と 3 に相当する．つまり，これらの基準が満たされていないのであれば，本来は Linked Data とは呼べない．そのため，データセットの提供方法が改善されることが望ましい．

基準 4 は 4 原則の 4 に相当するが，50 というリンク数には根拠がなく，外部リンクが極端に少ないデータセットを足切りするための数値以上の意味があるとは考えにくい．例えばヨコハマ・アート・LOD [*4] は現在 DBpedia Japanese へのリンクが 42 個のため，本家図では基準外になる．しかし，外部リンク数が多いから良いデータセットであるとは必ずしも言えない．

なお，本家図の基準に合致するだけで，データセットが本家図に採用されるわけではない．本家図に実際に採用されるためには，基準をクリアしていることを表明するための手続きが別に必要である．具体的には datahub [*5] に所定の方式

*3 http://opendefinition.org/
*4 http://fp.yafjp.org/yokohama_art_lod
*5 http://datahub.io

分類	データセット数	割合	トリプル数	割合	外部リンク数	割合
Industry	1	3.70%	87,983	0.02%	112	0.00%
Geographic	2	7.41%	6,398,759	1.70%	15,869	0.40%
Life Science	4	14.81%	140,510,938	37.39%	278,023	7.02%
Cross-domain	3	11.11%	108,000,143	28.74%	1,651,140	41.70%
Media	2	7.41%	33,137,619	8.82%	720,067	18.18%
Government	3	11.11%	5,415,553	1.44%	54,351	1.37%
Publication	11	40.74%	82,097,407 [1]	21.85%	1,238,166 [1]	31.27%
User generated content	1	3.70%	140,554	0.04%	1,994	0.05%
Total	27	100%	375,788,956 [2]	100%	3,959,722 [2]	100%

[1] CiNii 及び KAKEN は全体のデータセット取得が困難なため，その 2 個を除いた 9 データセット分の値である
[2] 1 同様，25 データセット分の値である

表 1: 分類別統計

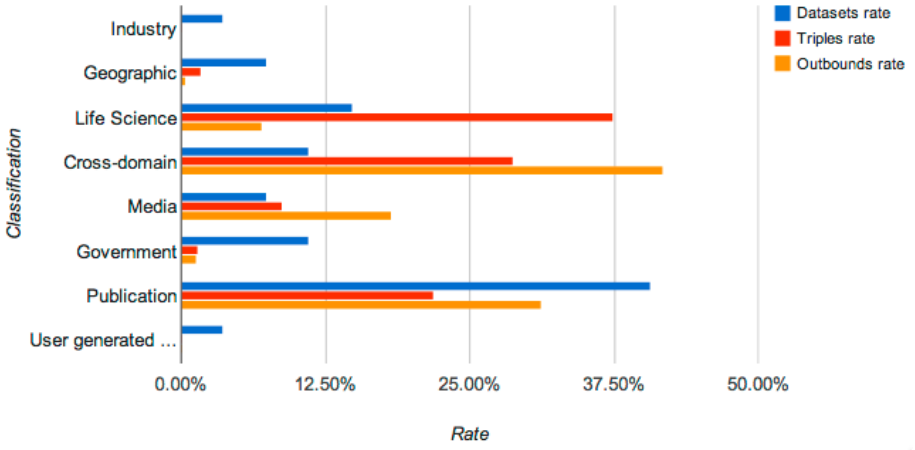

図 2: 分類別割合

でデータセットのカタログを記述した後に，Data Hub LOD Datasets*6 で個々の要件を確認する必要がある．日本語のデータセットでこの手続きを既に行っているのはわずかしかないが，今後本家図が更新されるときに採用されるためにも手続きを行なっておくことが望ましい．また，SPARQL エンドポイントがある場合は，この手続きによって SPARQL Endpoints Status*7 にも掲載されるようになる．

4. おわりに

本稿では日本語 Linked Data Cloud 図を通して，日本における Linked Data の現状と課題を紹介した．今後の課題としては，継続的に日本語 Linked Data Cloud 図を維持していくためには，手動で全ての調査を毎回行うには限界があるため，本家図のようなシステマチックな手段を検討する必要がある．

参考文献

[Ayers 08] Ayers, D, Völkel, M: Cool URIs for the Semantic Web, W3C Interest Group Note, http://www.w3.org/TR/cooluris/ (2008)

[Berners-Lee 06] Berners-Lee, T: Linked Data, http://www.w3.org/DesignIssues/LinkedData.html (2006)

[Cyganiak 11] Cyganiak, R, Jentzsch, A: The Linking Open Data cloud diagram, http://lod-cloud.net (2011)

[Fujii 13] Fujii, A, Egami, S, Shimizu, H: EDI support with LOD, 2013 Linked Data in Practice Workshop (2013)

[Kato 13] Kato, F, Takeda, H, Koide, S, Ohmukai, I: Building DBpedia Japanese and Linked Data Cloud in Japanese, 2013 Linked Data in Practice Workshop (2013)

[玉川 13] 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 Linked Open Data への取り組み, 第 27 回 人工知能学会全国大会論文集 (2013)

*6 http://validator.lod-cloud.net
*7 http://sparqles.okfn.org