

# Linked Dataと国内のデータ共有の動向

## A Report on Trends of Linked Data and Open Data in Japan

野本 昌子<sup>1\*</sup>

<sup>1</sup> ヤフー株式会社

<sup>1</sup> Yahoo Japan Corporation

**Abstract:** This short paper reports global trends of Linked Data and recent topics of Open Data in Japan. In ISWC(International Semantic Web Conference), the leading conference for research on Semantic Web, the number of papers on Linked Data has continued to increase for the past three years. Japan was ranked #31 in the Open Data Index (2015). At a session in WebDB Forum 2015, a major conference for research on DB and Web in Japan, issues one faces in making data open were revealed.

### 1 はじめに

本稿では Linked Data の最近の動向と国内のデータ共有の動向を紹介する。2015年10月にセマンティック・ウェブの最高峰の国際会議 ISWC<sup>1</sup> が米国で開催された。近年、同学会では Linked Data に関する発表が増加している。国内ではデータ公開が進んでいるが、Open Data Index という指標によれば日本のデータ公開の状況は世界の中では 31 位である。WebDB Forum2015<sup>2</sup> などではデータ共有の課題も議論された。

### 2 ISWC2015におけるLinked Dataの動向

2015年10月にセマンティック・ウェブの最高峰の国際会議である ISWC2015 が米国で開催された。最近の同学会の Linked Data に関する動向をプログラム構成、論文のキーワード別発表数、セッション別発表数などから概観する。

#### 2.1 プログラム構成

本学会の最近の主なプログラム構成を表 1 に示す ([1][2][3][4][5])。ISWC2015 から Data Sets and Ontologies という Track が設けられた。実際に採択された論文は 8 本に留まったが、投稿された論文は 35 本に及ん

だ。ISWC2016<sup>3</sup> でも Resources という Track が予定されているのはデータセットに関する論文の最近の増加傾向を反映したものと思われる。

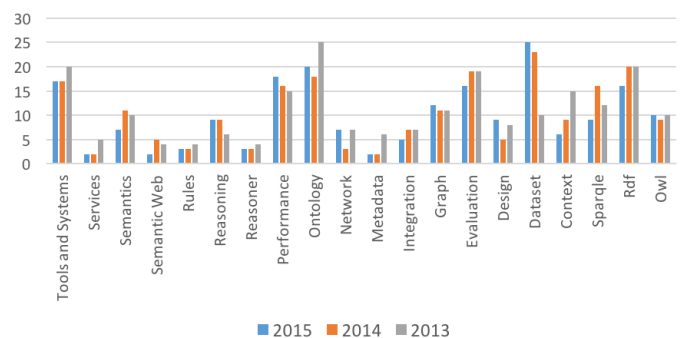


図 1: ISWC2013-2015 キーワード別発表数

#### 2.2 キーワード別発表数

また、学会の統計情報のサイト<sup>4</sup>では2002年から2015年までの論文のトレンドの推移を20のキーワードの分布(Streamgraph)で示している。このうち直近3年分のデータを切り出したものを図1に示す。データセットに関する発表がISWC2014で急増し、その後も増加傾向が続いている。

#### 2.3 セッション数とセッション別発表数

ISWC2015の本会議のセッションは Linked Data が 3 セッション、OBDA(Ontology Based Data Access)

\*連絡先: ヤフー株式会社

〒107-6211 港区赤坂 9-7-1

E-mail: mnomoto@yahoo-corp.jp

<sup>1</sup><http://iswc2015.semanticweb.org/>

<sup>2</sup><http://db-event.jp.org/webdbf2015/>

<sup>3</sup><http://iswc2016.semanticweb.org/>

<sup>4</sup><http://scientometrics.geog.ucsb.edu/iswc/>

表 1: ISWC のプログラム構成

ISWC2016 (2016.2.10 現在)	Track	採択率 (採択数/投稿数)			
	ISWC2015	ISWC2015	ISWC2014	ISWC2013	ISWC2012
Research Track	Research Track	0.221(38/172)	0.211(38/180)	0.214(45/210)	0.224(41/183)
	Empirical Studies and Experiments	0.292(7/24)	-	-	0.195(8/41)
In-Use	In-Use and Software	0.412(14/34)	In Use: 0.326(15/46)	In Use: 0.178(16/90)	In Use: 0.221(17/77)
Resources	Data Sets and Ontologies	0.229(8/35)	-	-	-
	Industry	-(20/ )	0.769(30/39)	-	0.632(12/19)
Posters and Demos	Posters and Demos	-(99/ )	0.776(121/156)	-	0.492(31/63)

表 2: ISWC2015 Linked Data セッショントピック例

医薬情報検索システム, Wikipedia Commons から知識抽出した DBpedia Commons の開発, 上位オントロジー (DOLCE) を用いた DBpedia のエラー検出, 薬物相互作用データ, SPARQL クエリログ, Linked Data の可視化, 大規模データでの評価環境構築, 都市に関するオープンデータ収集と欠損値の推定, LOD のキャッシュの更新方法, 関連エンティティランキング, エンティティナビゲーション, 教材の自動生成

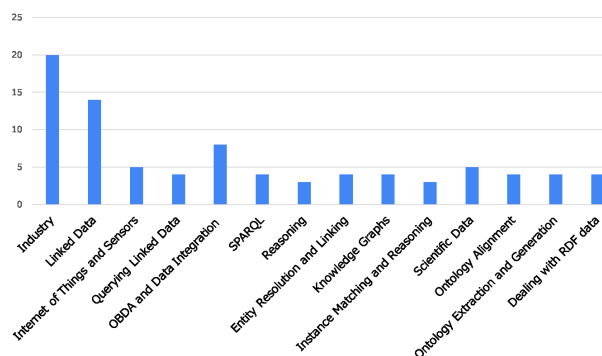


図 2: セッション別発表数

表 3: Linked Data 関連セッションの発表数の推移

開催年	2015	2014	2013
発表数	14	11	5

and Data Integration が 2 セッション, Industry 系は 3 セッションであった. Linked Data のセッションの論文のトピックは多岐にわたる (表 2).

今回のセッション別発表数を図 2 に示す. Linked Data は Industry の次に多く, 14 件であった.

Linked Data 関連セッションの発表数の過去 3 年間の推移を表 3 に示す. 2013 年は 5 件にとどまっていたが, 2014 年に計 11 件 (Linked Data & Data Quality: 5 件, Linked Data: 6 件) と急増し, 増加傾向が続いている. この Linked Data の発表数の増加は, 前述のキーワード別発表数の Dataset の発表数の増加傾向にも影響している可能性がある.

## 2.4 受賞論文

ISWC2015 の Best Research Paper が Linked Data の 実験環境を提供するプラットフォームに関する [6]

に与えられたことも注目に値する. 現在のセマンティック・ウェブ研究では DBpedia など一部のデータセットで評価をしているため評価結果の一般性が損なわれているが, 本来は Linked Data の多様性を考慮して評価をすべきだという問題認識に立ち, 異種の Linked Data を統一フォーマットに変換し, コマンドラインから実験に使用するデータの特徴を指定することで適切なデータセットで大規模な実験が簡単に行えるようにしたものである.

## 3 日本におけるデータ共有の動向

日本でのデータ公開の取り組みはアカデミックで先行して行われていたが, 近年の政府のオープンガバメントの取り組みを受けて地域情報のオープンデータが進み, またオープンデータに関する多くのコンテストが開催されている ([7][8]). また, Linked Open Data Initiative<sup>5</sup>, Open Knowledge<sup>6</sup> の日本グループである

<sup>5</sup><http://linkedopendata.jp/>

<sup>6</sup><http://okfn.org/>

Rank	Dataset	Breakdown	Location (URL)	Format	Info	Prev. (2014)	Score
1	Location datasets		<a href="http://www.post.japanpost.jp/z...">http://www.post.japanpost.jp/z...</a>	CSV		#1 100%	100%
8	Government Spending		n/a	n/a		#15 10%	10%
16	Legislation		<a href="http://law.e-gov.go.jp/cgi-bin...">http://law.e-gov.go.jp/cgi-bin...</a>	n/a		#12 75%	75%
20	Weather forecast		<a href="http://www.jma.go.jp/jp/week/">http://www.jma.go.jp/jp/week/</a>	n/a		n/a	55%
25	Government Budget		<a href="http://www.mof.go.jp/budget/re...">http://www.mof.go.jp/budget/re...</a>	Excel		#22 70%	70%
27	Pollutant Emissions		<a href="http://soramame.taiki.go.jp/">http://soramame.taiki.go.jp/</a>	n/a		#13 70%	55%
31	Land Ownership		<a href="http://www1.touki.or.jp/">http://www1.touki.or.jp/</a>	n/a		n/a	30%
33	National Statistics		<a href="http://www.e-stat.go.jp/SG1/es...">http://www.e-stat.go.jp/SG1/es...</a>	CSV, ...		#23 70%	70%
34	National Map		<a href="http://www.gsi.go.jp/kiban/ind...">http://www.gsi.go.jp/kiban/ind...</a>	GML		#20 70%	50%
36	Procurement tenders		<a href="http://www.kantei.go.jp/jp/kan...">http://www.kantei.go.jp/jp/kan...</a>	n/a		n/a	45%
46	Water Quality		n/a	n/a		n/a	10%
47	Company Register		<a href="http://www.touki.or.jp/">http://www.touki.or.jp/</a>	n/a		#57 30%	25%
87	Election Results		<a href="http://www.soumu.go.jp/senkyo/...">http://www.soumu.go.jp/senkyo/...</a>	Excel		#22 70%	0%

図 3: Open Data Index(2015) の日本の評価結果

表 4: データ提供者から見たデータ共有の課題

課題の分類	課題
情報共有/ 可視化	データ活用の知見 データ公開による効果, 公開前後の 状況変化の可視化 対象データの選定条件
情報伝達/ 意識合わせ	現場の課題 データ作成の背景 データの選定方法, 組み合わせ 使いやすいデータ形式
プライバシー 効率化	個人情報の配慮, プライバシー保護 データ加工 (クリーニング・整形等) の工数削減 公開準備タスク (契約など) の簡素化, 効率化
即時性	フィードバックのタイムラグ等

オープン・ナレッジ・ファウンデーション・ジャパン (OKFJ)<sup>7</sup> 等で LOD(Linked Open Data) の普及促進やデータ開発, 利用の支援をする活動も行われている。

### 3.1 Open Data Index によるデータ共有状況の評価

世界の中での日本のオープンデータの進捗状況の一つの目安として, government data については Open

<sup>7</sup><http://okfn.jp/>

Data Index<sup>8</sup> という指標がある。これは前述の Open Knowledge が世界の国を対象に複数分野でオープンデータの入手しやすさと扱いやすさを評価したものである。

2015 年は 122 の国を対象に 13 分野で評価が行われ, 1 位は台湾, 日本は 31 位であった。前記 13 分野とは Location datasets, Government Spending, Legislation, Weather forecast, Government Budget, Pollutant Emissions, Land Ownership, National Statistics, National Map, Procurement tenders, Water Quality, Company Register, Election Results である。Open Data Index(2015) の日本に関する評価結果を図 3 に示す。上記 13 分野のうち, 日本でもっともオープンデータ化が進んでいるのは Location datasets(100%), 続いて Legislation(75%), Government Budget 及び National Statistics(70%) である。逆に遅れているのは Election Results(0%), Government Spending 及び Water Quality(10%), Company Register(25%) などである。

なお, 2014 年は Weather forecast, Land Ownership, Procurement tenders, Water Quality を除き Transport Timetables を加えた 10 分野での評価で 1 位は英国, 日本は 19 位であった。

### 3.2 データ共有の課題

昨年 11 月に開催された WebDBForum2015 では特別セッション「産学間のデータセット共有の意義, 課題と将来の展望」においてデータの利用者, 提供者, デー

<sup>8</sup><http://index.okfn.org/>

タリポジトリの責任者の各々の立場からの発表とディスカッションが行われた。

国立情報学研究所では情報学研究データリポジトリ<sup>9</sup>で研究用のデータを公開している。企業の提供したデータとしては従来からの楽天データセット、Yahoo!知恵袋データ、ニコニコデータセットに加えて、2014年はリクルートデータ、2015年にはクックパッドデータ、HOME'Sデータセットが追加された。発表時点で、約440研究室にデータを提供し、そのデータを用いて約350論文が発表されているという([9])。

データの利用側からは正解データの提供方法に関する課題が指摘され、データの内容によりNDA配布とリポジトリでの配布を使い分けることが提案された。

一方、データの提供者である企業側からはデータ公開による効果としてデータ利用組織の大幅増や研究の広がりなど、データ公開のメリットとともに、公開に関する課題も指摘された([10][11][12])。主な課題をまとめたものを表4に示す。

[10]では研究成果が(サービスの)ユーザに価値となって届くことが今後の「TODO」として挙げられていた。公開に関するタスクを整理、軽量化して、全体のサイクルを早く回し、タイムリーな改善につなげるのが重要と思われる。また、データ公開に関する知見の共有と、データの提供者と利用者のように立場の異なる関係者が継続的に連携して、意思疎通や相互理解をスムーズに進められるようにするための工夫が必要である。

## 4 おわりに

本稿ではLinked Dataと国内のデータ共有の最近の動向を紹介した。

セマンティック・ウェブの最高峰の国際会議ISWCでは近年Linked Dataに関する発表が増加している。

一方、データ公開についての世界の中での日本の位置付けはOpen Data Indexという指標によれば2015年は31位であった。国内ではデータ公開が進む一方、データ共有の課題も多い。今後データ共有が進むためには、データ公開に関する知見の共有や立場の異なる関係者間の相互理解を進めるための工夫が必要である。また、公開に関するタスクの整理、軽量化により、全体のサイクルを早く回してタイムリーな改善につながるようにすることが重要と思われる。

## 参考文献

- [1] Steffen Staab: ISWC2015 Opening Session, <http://www.slideshare.net/steffenstaab/iswc2015-opening-session> (2015)
- [2] Carole Goble, Abraham Bernstein: Opening and 10 years award presentation, [http://videlectures.net/iswc2014\\_goble\\_bernstein\\_opening\\_award/](http://videlectures.net/iswc2014_goble_bernstein_opening_award/) (2014)
- [3] 川村隆浩他: 国際会議 ISWC2014 参加報告, <http://sigsw.org/papers/SIG-SWO-035/SIG-SWO-035-11.pdf> (2014)
- [4] Alani, H., et al. (Eds.): The Semantic Web - ISWC 2013, <http://www.springer.com/jp/book/9783642413377> (2013)
- [5] Abraham Bemsterin, et al.: ISWC2012 Opening Session, [http://videlectures.net/iswc2012\\_bernstein\\_opening/](http://videlectures.net/iswc2012_bernstein_opening/) (2012)
- [6] Laurens Rietveld, et al.: LOD Lab: Experiments at LOD Scale, <http://laurensrietveld.nl/pdf/lodlab.pdf> (2015)
- [7] 古崎晃司: Linked Open Data チャレンジ Japan を中心とした LOD の事例紹介, <http://www.slideshare.net/KoujiKozaki/linked-open-data-japanlod> (2013)
- [8] 古崎晃司: LOD(Linked Open Data) の動向と今後の展望, <http://www.slideshare.net/KoujiKozaki/20150618-lod-kozakifinal> (2015)
- [9] 株式会社ネクスト Engineer Blog, <http://nextdeveloper.hatenablog.com/entry/2015/12/04/142050> (2015)
- [10] 原島純: クックパッドデータの公開, その舞台裏, <http://speakerdeck.com/junharashima/webdb-forum-2015> (2015)
- [11] 清田陽司: 「HOME'S データセット」提供開始の背景 ~産学間データ共有の課題~, <http://www.slideshare.net/ykiyota/homes-55449185> (2015)
- [12] 櫻井一貴: 「リクルートデータセット」~公開までの道のりとこれから~, <https://www.slideshare.net/recruitcojp/ss-55685537> (2015)

<sup>9</sup><http://www.nii.ac.jp/dsc/idr/datalist.html>