1 リンクト・オープン・データの原理原則と最近の進歩

応般

武田英明（国立情報学研究所）

LODとは何か？

リンクト・オープン・データ（LOD）とは，セマンティックWeb分野で開発・標準化がなされてきた技術による，Web上のデータを公開・利用する方式あるいは公開されたデータセットのことを指す．これまでのWebは「文書のWeb」であり，文書を相互にリンクしてネットワークを作っていたのに対して，LODは同様のネットワークをデータの間で作るため，「データのWeb」と呼ばれる☆1．LODはWebと同じようにグローバルに共有されるデータ空間を作り出す．すなわち，インターネット上に公開されたデータであれば，それがどこのだれが管理しているデータセットに含まれているかを意識することなく，アクセスしたり，リンクしたりすることができる．個々のデータセットはお互いにリンクし合うことで，1つのグローバルなデータ空間の一部となる．

LODはデータの公開・共有手段として広く認知され，いまやさまざまな分野で実用的なデータ公開に用いられている．本特集では特に各分野での事例に焦点を当てるが，本稿ではLOD全体のこれまでの経緯や現在の状況についてまとめる．

LODの発展の経緯

LODはセマンティックWebの発展の延長線上にある．セマンティックWebからLODが生まれた理由には，セマンティックWeb技術の進展もあるが，Web自身の発展によるところが大きい．セマンティックWebによってLODの技術的な仕組みは提供されたものの，それだけでは当初大きな普及には至らなかった．その理由の1つとして普通のWebページをセマンティックWebに対応させるためには適切なメタデータを付与するという労力が必要な一方，一定規模のセマンティックWeb対応の情報集積がないとメリットが得られにくく，大きな広がりにならなかった．

一方，Webが着実に社会の基盤としてあらゆる情報を含むようになるにつれて，販売品のカタログや図書館の目録など，既存のサービス情報やデータベース中の情報をHTMLとして提供するWebページが急速に増えてきた．この場合，Webページの情報は人間向けにHTMLを使って記述されたものの，元の情報の持つ構造は必ずしも適切に表現されなかった．そのような状況において，セマンティックWebの技術が注目されることになった．

データベースから生成される情報のWebページのように，もともと構造を持つデータの場合には，メタデータの付与はプログラムが情報を変換するときに行うことができるので，その生成時の人的負荷は少なく，メタデータを付与した大量のWebページを生成できる．さらに一歩進んで，そもそも人間可読のWebページとは別に，もともとの意味構造を保持したまま機械可読なデータを公開すれば，プログラムは容易にそのデータを利用することができる．

人間可読から機械可読データへの動きとして2007年から始まったものがLinking Open Data Projectである．このプロジェクトで構築されたDBpediaが大きな契機となって発展が始まり，現在では世界中で各種各領域のデータがLOD化され，

☆1 LODを中心トピックスとする最初のworkshop（LDOW2008, Linked Data on the Web）[1]において，Web of DataとしてのLODが明確に謳われている．

DBpedia を中心にそれらが大量の情報が互いにリンクされ公開されるようになった.

LOD の基本

LOD は技術的にはセマンティック Web のそれと変わらない．セマンティック Web が文章等の非構造あるいは半構造情報を対象とするのに対して，LOD は構造的なデータを対象とする．このため，セマンティック Web では情報にメタデータを付与するという形をとったのに対して，LOD では構造が意味情報であるので，データ自身が RDF（Resource Description Framework）[☆2] を用いて記述される．RDF では，< 主語，述語，目的語 > という 3 つ組みで情報を表現する．3 つ組みの集合は全体として主語と目的語をノードとし述語をアークとするグラフ構造を形成するが，このグラフは RDF グラフと呼ばれる．これらグラフの構成要素は URI [☆3] あるいはリテラル（文字列や数値）で記述される．LOD では，データセット中の個別の事物や事象にユニークな URI を与える．そして RDF を用いて URI が表示するものの間の関係を記述する．URI は Web 上でユニークに識別可能なので，そのデータがどのデータセットのものかを気にすることなく，URI だけで事物・事象の関係や性質を引用でき，グローバルに共有するデータ空間を構成することができる.

データが持つ意味構造は別途スキーマとして定義するが，LOD ではリンクの持つ意味を RDF Schema（RDFS）[☆4] や OWL（Web Ontology Language）[☆5] により，クラスとプロパティの組合せとして定義する．クラスは事物や事象の概念を示し，個々の具体的事実の型を表す．プロパティは個々の事実の関係を表すものであるが，RDFS や OWL ではクラスが持つ属性の型やクラス間の関係の型を指定して，個別の事物や事象はこれらのクラスのインスタンスとして記述される．このようなスキーマ定義はデータセットごとに用意する必要はなく，すでにほかで定義されたスキーマが使えるときは，それを使えばよい.

RDFS は RDF の拡張であるので，クラスやプロパティの定義にも URI を用いる．URI と RDF を使うことで個別の事物や事象のみならず，それらのスキーマも共有してデータを相互につなげることができる．これは新しいデータの世界である.

Tim Berners-Lee はこの LOD を普及させるために以下の 4 つの原則を提唱した[☆6].

1. URI を使って事物を名前付けしよう
2. 名前の参照が HTTP URI でできるようにしよう
3. URI を参照したときに関連情報が手に入るようにしよう
4. 外部へのリンクも含めよう

これらの原則に基づくことによって，データが相互につながるようになる．たとえば http URL を用いることで，個々のデータは通常の Web 技術を用いてアクセス可能となる．http URL をデータに付与するのみならず，実際に Web 上でアクセス可能にすることを参照解決可能という．コンテント・ネゴシエーションは HTML ヘッダ部分の情報によって文字情報を与えたり，フォーマットを指定したりするためのものであるが，この技術を用いれば，人間用には人間可読なデータ表示を与え，機械用には機械可読なデータそのものを与えることが可能になる．また，データセットが外部リンクを持つことで Web サーフィンするように，人間や機械がデータサーフィンすることができるようになる.

さらにデータを自由に検索できれば，データをより活用できる．RDF ストアとは，RDF データを蓄積保存し，検索可能にしたもので，現在いくつかの無料で使えるソフトウェアや商用のシステムがある．RDF ストアに対する検索用のクエリ言語が SPARQL [☆7] で，Web を経由して外部から検索可能にしたサイトを SPARQL エンドポイントと呼ぶ.

☆2 http://www.w3.org/TR/rdf11-primer/
☆3 URL の一般化，なお国際化 URI である IRI が本来適切であるが以下では簡便のため URI と呼称.
☆4 http://www.w3.org/TR/rdf-schema/
☆5 http://www.w3.org/TR/owl2-overview/
☆6 http://www.w3.org/DesignIssues/LinkedData.html
☆7 http://www.w3.org/TR/sparql11-overview/

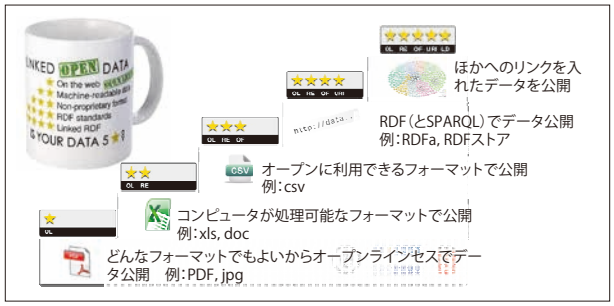

図 -1　オープンデータの 5 つ星

RDF ストアはいわゆるスキーマレスの NoSQL の一種と見ることができる．RDF と RDF ストアを用いて，データベースのサイロ化の問題を解決することができる．バイオサイエンス系では従来から多くのデータベースがあるが，これらのデータベース中のデータを一元的に統一的なアクセス方法で取得できるようにするために LOD を利用したデータベースの統合化が進められてきた．詳細は本特集の「生命科学分野における LOD の構築と利用」（山本）を参照されたい．

LOD とオープンデータ

リンクト・オープン・データとはリンクされたオープンなデータのことではあるが，リンクトデータすなわちオープンデータではない．Tim Berners-Lee は LOD 構築の視点からオープンデータの 5 つ星（Five Star Open Data）という階層を提案した☆8．図 -1 にこれを図示する．

単に Web 上にデータを公開するだけではオープンデータではない．公開データにオープンライセンス☆9 を付与して初めて，星 1 のオープンデータと見なす．この段階では公開のデータ形式は問わないが，次に " 機械可読な " 形式で公開することを星 2 と見なす．さらにこのデータ形式が独占所有権のないオープンな形式であるとき，星 3 とする．ここまでが " 一般の " オープンデータのレベルである．さらにデータ形式が RDF であるときを星 4 と見なす．加えて外部へのリンクを含むような RDF で書いたときを星 5 と見なす．これがまさに LOD である．この最後の段階でデータは相互につながりあい，グローバルなデータ空間の一部として活用可能になる．

なお，オープンデータは Web 上に散らばっているため，これを集めると利便性が上がる．多種多様なオープンデータを一同に集めてそのメタデータにより検索しやすくしたものをオープンデータ・カタログサイトと呼ぶ．事実上の標準として普及しているプラットフォームに CKAN がある．CKAN では DCAT というメタデータ・スキーマが使われている．たとえば日本政府によるオープンデータ・カタログサイト☆10 も CKAN で作られている．

LOD の実際

データセットの公開としては，2007 年の DBpedia の公開が大きな契機となった．DBpedia は多様な分野のデータを含み，ほかのデータセットとリンクが容易であるため，DBpedia を中心に LOD のネットワークが自然発生的にできた．これを LOD クラウドと呼ぶ．国内においても 2012 年に Wikipedia 日本語版から生成された DBpedia Japanese が公開され，国内の LOD ネットワークの中心として機能している．本特集の「地理空間情報と LOD」（松澤）図 -3-（a）および図 -3-（b）を参照されたい．

スキーマに関しては多くのデータセットにおいて文書のメタデータを記述するための Dublin Core☆11 や，人に関するメタデータを記述するための FOAF（Friend-Of-A-Friend）☆12 が使われている．このほか分野特有のものもある．SKOS（Simple Knowledge Organization System）☆13 はもともと図書館情報における分類体系や件名標目（Library of Congress Subject Headings，LCSH）表などの準形式的な知識組

☆8 http://5stardata.info/ja/

☆9 オープンライセンスについては過去の特集記事「オープンデータ活用 [2)]」を参照されたい．

☆10 http://www.data.go.jp/data/dataset

☆11 http://www.kanzaki.com/docs/sw/dublin-core.html

☆12 http://www.kanzaki.com/docs/sw/foaf.html

☆13 https://www.w3.org/TR/skos-primer/

織化体系に基づいたスキーマであるが，RDFS の厳密な意味論に従わなくてもよいため，現在は RDFS に代わる分類記述のためによく使われている．

実際に既存のデータを LOD 化しようとすると，関係やクラスの記述に何を用いたらよいかが問題となる．その場合，LOD 本来の目的のためには皆が共通に用いている語彙を使うのが望ましい．LOV（Linked Open Vocabulary）というサイト ☆14 や RDF で用いられる名前空間をまとめたサイト prefix.cc ☆15 を検索することで，不必要に新しいスキーマを生成せず，なるべくスキーマを共有するというセマンティック Web の基本思想が実践されている．

最近の動向

LOD に関する技術は 2012 年から 2014 年までに一通りの技術標準（RDF, Turtle ☆16, JSON-LD ☆17, OWL, SPARQL 等）が制定されてきた．現在はこれらの技術を使ったシステム，アプリケーション，サービスが作られ，アーリーアダプタ段階にあるといえる．国内では 2012 年頃までは学術的関心が主であったが，2013 年頃よりオープンデータ活動の盛り上がりに伴い，オープンデータの次世代技術として広く関心を集めるようになった．特筆すべきは商用を含むいくつかの LOD 公開支援サービスが登場して，技術的知識がなくても保有するデータを LOD として公開できる環境が整ってきたことである．LinkData.org ☆18 は早くから簡便な LOD 化ツールを提供してきたが，jig.jp 社ではオープンデータプラットフォームという有料サービス ☆19 を 2014 年 6 月より開始し，地方自治体などが採用している．このほか，Datashelf（インフォラウンジ社）☆20 などもある．

一方，これらのイノベータやアーリーアダプタの経験から不足する機能が明らかになり，それが新たな技術要素が標準にフィードバックされ，第 2 弾の標準化のプロセスに入っている．以下では代表的な技術分野ごとに最近の動向も交えて，本特集記事との関連を紹介する．

● LOD 実践ベストプラクティス

Linked Data Platform（LDP）は http プロトコルによる読み書き可能な Linked Data のアーキテクチャを規定するもので，2015 年 2 月に W3C 勧告となった ☆21．従来から適切な URI の書き方としてクール URI ☆22 が提唱されていたが，LDP は LOD に焦点を合わせて，put/get や HTML ヘッダの書き方まで発展させたものである．

CSV のような表形式のデータを LOD にするということは実際によく行われることである．CSV on the Web は表形式のデータを LOD で扱いやすくするための標準で，CSV ファイルの記法，メタデータ語彙，JSON 変換方法，RDF 変換方法など規定され，2015 年 12 月 W3C 勧告 ☆23 となっている．

RDF は当初からドメイン横断的なデータ連携を目的としていたが，RDF が普及するにつれてデータの発生からアプリまでの RDF ツール横断的な可用性や相互運用性が意識されるようになってきた．RDF Data Shapes ☆24 は，データ流通を目的に RDF グラフの形（Shape）データの検証やインタフェース仕様において必要となる RDF に対する構造的な制約を規定するもので，現在はまだ議論中である．

アーリーアダプタとしての適用事例はクロスドメイン関係，図書館関係，バイオサイエンス関係，政府データ関係が先導的である．

● クロスドメイン関係

クロスドメイン関係とは辞書や事典のように分野を超えて横断的に使われるデータを指す．この

☆14 http://lov.okfn.org/dataset/lov/
☆15 https://prefix.cc/
☆16 http://www.w3.org/TR/turtle/，RDF 記述のための簡潔な書法
☆17 http://www.w3.org/TR/json-ld/，JSON による RDF 記述の書法
☆18 http://linkdata.org/
☆19 http://odp.jig.jp/
☆20 http://datashelf.jp/
☆21 http://www.w3.org/TR/ldp/
☆22 http://www.kanzaki.com/docs/Style/URI
☆23 http://www.w3.org/TR/2015/REC-tabular-data-model-20151217/
☆24 http://www.w3.org/TR/shacl-ucr/

関係では Wikipedia をデータ化した DBpedia が著名であり，LOD を先導してきた．2015 年には Wikipedia で参照される画像のデータセットである Wikimedia commons も LOD 化された（DBpedia Commons）☆25．また，社会的・公共的な情報を文章ではなくデータとして Wikipedia 同様に共同作業で構築していこうという Wikidata プロジェクト☆26 が進行中であり，これも LOD になっている．国内では 2012 年より日本語 Wikipedia を LOD 化した DBpedia Japanese が公開されている．

● 図書館・博物館関係

図書館における書誌および典拠はもともと公開・共有されるものであり，かつ構造化されており，さらには相互参照されるものであったので，LOD と相性がよい．このため各国の中央図書館が積極的にデータを LOD として公開している．2009 年 4 月の米国議会図書館による件名標目の公開を皮切りとして，名称典拠や分類表，各種コードなど約 40 種類の情報を LOD として公開している．欧州では英国，ドイツ，スイス，フランス，スペイン等の国立図書館が順次データ公開を始めている．国内の図書館関係ではすでに国会図書館の典拠データ（本特集の「出版物に関するメタデータと国際書誌コントロール」（橋詰）を参照），国立情報学研究所の論文検索サービス CiNii の RDF 化☆27 などが行われていたが，さらに 2015 年より科学技術推進機構の J-Global の RDF 化☆28 もスタートしている．

また，博物館関係でも同様に公開が進んでおり，大英博物館では 2011 年 9 月に所蔵コレクションのデータを LOD で公開を始めた☆29．ヨーロッパの図書館，博物館等のデータを収集・公開する Europeana☆30 でも順次ライセンスがオープンになったものを LOD として公開している．またゲッティ研究所☆31 は 2014 年から 2015 年にかけて美術関係のシソーラス（Art & Architecture Thesaurus，AAT）と地名典拠（Getty Thesaurus of Geographic Names，TGN），芸術家典拠（Union List of Artist Names，ULAN）を順次 LOD 化して公開した☆32．

● バイオサイエンス関係

バイオサイエンス分野ではもともと多様なデータを大量に利用しており，その整理のためにオントロジー（概念の体系化）が利用されており，LOD との親和性が高い．このため，米国国立医学図書館（NLM）が公開する生命科学用語集 MeSH（Medical Subject Headings）☆33 を始めとして，多くのデータが LOD 化されている．前述のように，バイオサイエンス分野では多数のデータベースがあるため，それらを横断的につなぐのに LOD は有効であり，たとえば bio2rdf プロジェクト☆34 では 19 のデータセットを RDF 化してつないでいる．本特集「生命科学分野における LOD の構築と利用」（山本）も参考にされたい．

● 地理情報関係

LOD アプリを作るとき，各種のデータを地図の上にマップして表示すると格段と魅力が増す．地理空間情報システムは一種のアプリケーション・プラットフォームと見ることもできるが，LOD 普及に伴って，地理関係データのハブとなる地理識別子そのものの LOD 化の要求も増してきている．国際的にはこの動きを geonames.org☆35 が推進しているが，国内では 2015 年 5 月より地名の URI 基盤として geonames.jp がスタートしている．その詳細は本特集の「地理空間情報と LOD」（松澤）を参照されたい．

☆25 http://commons.dbpedia.org/
☆26 https://www.wikidata.org/
☆27 https://support.nii.ac.jp/ja/cia/api/a_rdf
☆28 https://stirdf.jglobal.jst.go.jp/
☆29 http://collection.britishmuseum.org/
☆30 http://www.europeana.eu/portal/
☆31 http://www.getty.edu/research/
☆32 http://www.getty.edu/research/tools/vocabularies/lod/
☆33 https://www.nlm.nih.gov/mesh/
☆34 https://github.com/bio2rdf/bio2rdf-scripts/wiki
☆35 http://www.geonames.org/

● 政府・地方自治体関係と共通語彙基盤

政府データ関係では政府データのオープンデータ化の際の手段としての LOD が浸透しつつある．特に英国では各種のデータが LOD として公開されている．たとえば陸地測量部（Ordnance Survey）では地理関係の大規模データセットを 2010 年より LOD により公開している．また英国ではコミュニティ・地方自治省（Department for Communities and Local Government，DCLG）が多様な地方行政データを LOD として公開しており，現時点で 215 データセットが登録されている．

英国および米国のオープンデータは先進的であるが，日本政府においても前述のオープンデータ・カタログサイトに見られるように，政府の保持するデータは原則オープンデータとして公開することになっている．さらに，従来の政府標準利用規約（第 1.0 版，2014 年）を 2015 年 12 月に第 2.0 版として改定し，CC BY 4.0 ☆36 互換とされた ☆37．これにより地方自治体含め政府関連データのオープンデータ化はさらに進むものと期待される．

地方自治体のオープンデータの LOD としての公開も増えてきているが，各所でバラバラに用いられている用語をそのまま LOD 化するのでは，その LOD に期待される効果は大きく減じられる．経済産業省および情報処理推進機構（IPA）が推進する共通語彙基盤では，基本的な語彙を定義して RDF 形式および XML 形式で提供している．その詳細は本特集「政府が推進する社会のデータ共有環境の整備」（平本）を参照されたい．北海道森町のように共通語彙基盤のコア語彙を用いて LinkData.org 上で公開を試みる例も出てきた．また，滋賀県大津市のように，びわ湖花火大会のデータを独自に LOD として公開する例も見られる．ハッカソンとはもともとは地方自治体や公共団体の抱える問題を地域のコミュニティによる解決を目的に，ソフトウェアエンジニアが集まって集中的に 1 ～数日でプレプロトタイプを作り上げるという運動のことであるが，同種のことがオープンデータの利用方法や，LOD 作成とアプリ開発でも行われるようになった．関西では，大阪市や和歌山県等が公開しているオープンデータを有志が LOD 化して再公開している．またこれらのデータを使ったアイデアソン，ハッカソンも数多く開催されている．本特集の「シビックテックと LOD」（古崎）を参考にされたい．今後，地方自治体を含む政府関係のより多くのデータがオープンデータのみならず LOD としてますます公開されることが期待される．

☆36 CC BY 4.0 とは Creative Commons が推進するライセンスの 1 つで，適切なクレジットを表示する限り，複製，配布，利用ができるライセンスである，https://creativecommons.org/licenses/by/4.0/

☆37 政府標準利用規約第 2.0 版自身は CC BY 4.0 とは別に規定されるライセンスであるが，その規定の中に CC BY 4.0 互換であると記されており，CC BY 4.0 と読み替えてよい

参考文献

1）Bizer, C., Heath, T., Idehen, K. and Berners-Lee, T. : Linked Data on the Web（LDOW2008），In Proceedings of the 17th International Conference on World Wide Web（WWW 2008), ACM, New York, NY, USA, pp.1265-1266, DOI:10.1145/1367497.1367760
2）特集「オープンデータ活用」，情報処理，Vol.54, No.12, pp.1202-1247 (Dec. 2013).

（2016 年 5 月 3 日受付）

武田英明（正会員）■ takeda@nii.ac.jp

国立情報学研究所・情報学プリンシプル研究系・教授．総合研究大学院大学・教授．1991 年東京大学工学系研究科博士課程修了．工学博士．1992 ～ 1993 年ノルウェー工科大学，1993 年～ 2000 年奈良先端科学技術大学院大学を経て現職．