オープンデータと Linked Open Data

基般

大向一輝（国立情報学研究所）

オープンデータの技術面

Web を通じたデータの積極的な公開・共有を目指すオープンデータを進めるにあたっては，ライセンスの選択やデータが作られる現場のワークフロー設計といった制度面における課題とともに，公開されるデータのフォーマットやアクセス方法などの技術面での検討が必要である．実際にオープンデータを利用するユーザの観点からは，入手が容易であり，処理のしやすいデータほど活用されやすい傾向にあると思われる．Web 上に散在する多様なデータに対して統一的な手段でアクセスすることができ，そのデータが共通のルールに基づいて記述されているような環境の構築は，オープンデータの普及にとって重要な課題である．これに対して，文書の公開・共有手段として成功を収めた Web の技術的方法論をデータに適用する Linked Open Data（LOD）が注目されている．本稿ではオープンデータを支える技術としての Linked Open Data について述べ，今後の展望について議論する．

オープンデータと Linked Data の交点

オープンデータは単にデータを公開することではなく，二次利用や商用利用が認められた形でのデータ提供を指す．データにオープンライセンスを付与することは，第三者による情報の利活用を実現するための最低条件であるといえる．一方，オープンライセンスが付与されてさえいれば自動的にデータの利用が活性化されるとは限らない．現実には多くの Web サイトにおける情報公開が，コピー＆ペーストが不可能な画像情報や，サイトごとにマークアップの仕方が大きく異なる HTML 文書によって行われており，実際にデータが利用される機会は多くない．一部の先進的な Web サービスでは Web API が提供されており，XML や JSON といった構造化フォーマットでデータを得ることができるものの，サービス間で記法が統一されていないため，利用に際しては開発者が個別に対応する必要がある．

国内でデータ形式に関する議論が広まったきっかけは 2011 年 3 月の東日本大震災である．震災直後に電力不足の懸念が生じた際に，電力会社から電力供給量に関する情報が画像のみで提供されたことに端を発し，開発者らが再加工の可能なフォーマットでの配布を要求した結果，CSV 形式での提供が行われることとなった☆1．その数日後には電力供給状況を可視化するアプリケーションが多数開発され，データ形式の重要性が見直される契機となった☆2．同月末には経済産業省が日本経済団体連合会に対し，震災関連情報のデータ形式を自動処理に適した方法で提供するよう依頼した☆3．

オープンデータにおけるデータ形式の重要性を示した文書として，Web の提案者である Tim Berners-Lee による 5 star Open Data がある．5 star Open Data ではオープンデータが満たすべき条件を 5 段階のステップとして表現している（**図 -1**）．

最初のステップとして，オープンライセンスを付与して Web 上に公開されたデータには 1 つ星が与えられる（OL : Open License）．1 つ星のデータに

☆1 http://www.tepco.co.jp/forecast/index-j.html
☆2 http://itpro.nikkeibp.co.jp/article/NEWS/20110325/358756/
☆3 http://www.meti.go.jp/policy/mono_info_service/joho/other/2011/0330.html

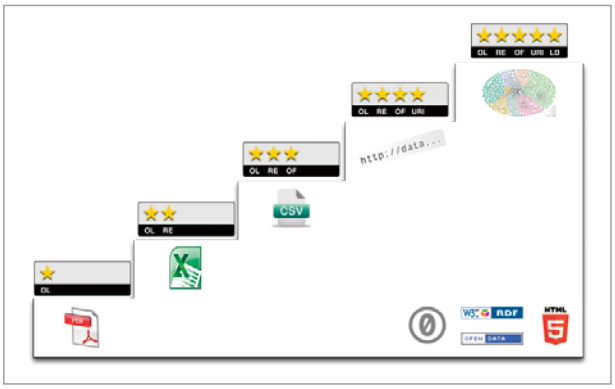

図 -1　5 star Open Data（http://5stardata.info）

おいてはフォーマットや形式はどのようなものでもよい．例としては画像化された PDF が挙げられている．次に，機械可読で再利用可能なデータには 2 つ星が与えられる（RE : Reusable）．表形式のデータであれば Microsoft Excel 形式で公開されたファイルなどがこれにあたる．非独占的なフォーマットで公開されているデータには 3 つ星が与えられる（OF : Open Format）．3 つ星の例としては CSV や XML が挙げられている．なお，.docx や .xlsx などの拡張子を持つ Office Open XML（OOXML）は ISO/IEC で標準化されているため 3 つ星に相当する．4 つ星のデータは，URI を用いて個々のデータを表現することで外部からのリンクが可能なものを指す（URI : Uniform Resource Identifier）．さらに，5 つ星のデータでは他のデータへのリンクが必須となる（LD : Linked Data）．

このステップの中で，4 つ星ならびに 5 つ星は Linked Data の概念に基づくデータ表現である．Linked Data はセマンティック Web 技術の応用として，データの意味論を RDF（Resource Description Framework）・RDF スキーマ・OWL（Web Ontology Language）を用いて記述する．Linked Data については文献 1）をはじめとして本誌 2011 年 3 月号の特集「リンクするデータ」に詳しい．ここでは Tim Berners-Lee による Linked Data の 4 原則[2)]を確認するにとどめる．

1. あらゆる事物の識別子に URI を使用する
2. 識別子には HTTP URI を使用する
3. URI にアクセスすると事物に関する構造化データが得られる
4. 構造化データには他の事物へのリンクを含む

これらの仕組みを用いることで，HTML とハイパーリンクによる「文書の Web（Web of Documents）」と同様の「データの Web（Web of Data）」を構築することが Linked Data の目標である．

なお，Linked Data 自体はデータ形式に関する技術的方法論であり，対象とするデータがオープンであるかどうかはこだわらない．その意味でオープンデータと Linked Data は直交した概念であり，5 star Open Data において 4 つ星，5 つ星に Linked Data の要素が含まれているのは適切でないとの批判もある．しかしながら，情報源ごとにデータ形式が異なっていることが情報の利活用を妨げているのは事実であり，データ形式の標準化とリンクによって新たな情報空間を構築するというビジョンを支持する声も大きい．データ同士をつないで再利用性を高めるという Linked Data の考え方は，オープンガバメントの旗手ともいえる米国連邦政府の data.gov ならびに英国政府の data.gov.uk においても強く支持されている．本稿では，この Linked Data の技術に基づくオープンデータ，すなわち Linked Open Data（LOD）について述べていく．

Linked Open Data の現在

LOD に対応する Web サイトは年々増加している．本稿ではこれらのサイトにて提供されているデータ群をデータセットと呼ぶ．データセットの総数や全体的な傾向は LOD cloud diagram で見ることができる（**図 -2**）．この図では，個々のノードがデータセットに対応し，エッジがデータ同士のリンク関係に相当する．また，データセットは Media や Geographic，Government など 7 種類に分類され，色分けされている．

LOD cloud diagram には，英国 Open Knowledge Foundation が運営するオープンデータのカタログサイト Datahub[☆4] に登録されているデータセット

☆4 http://datahub.io

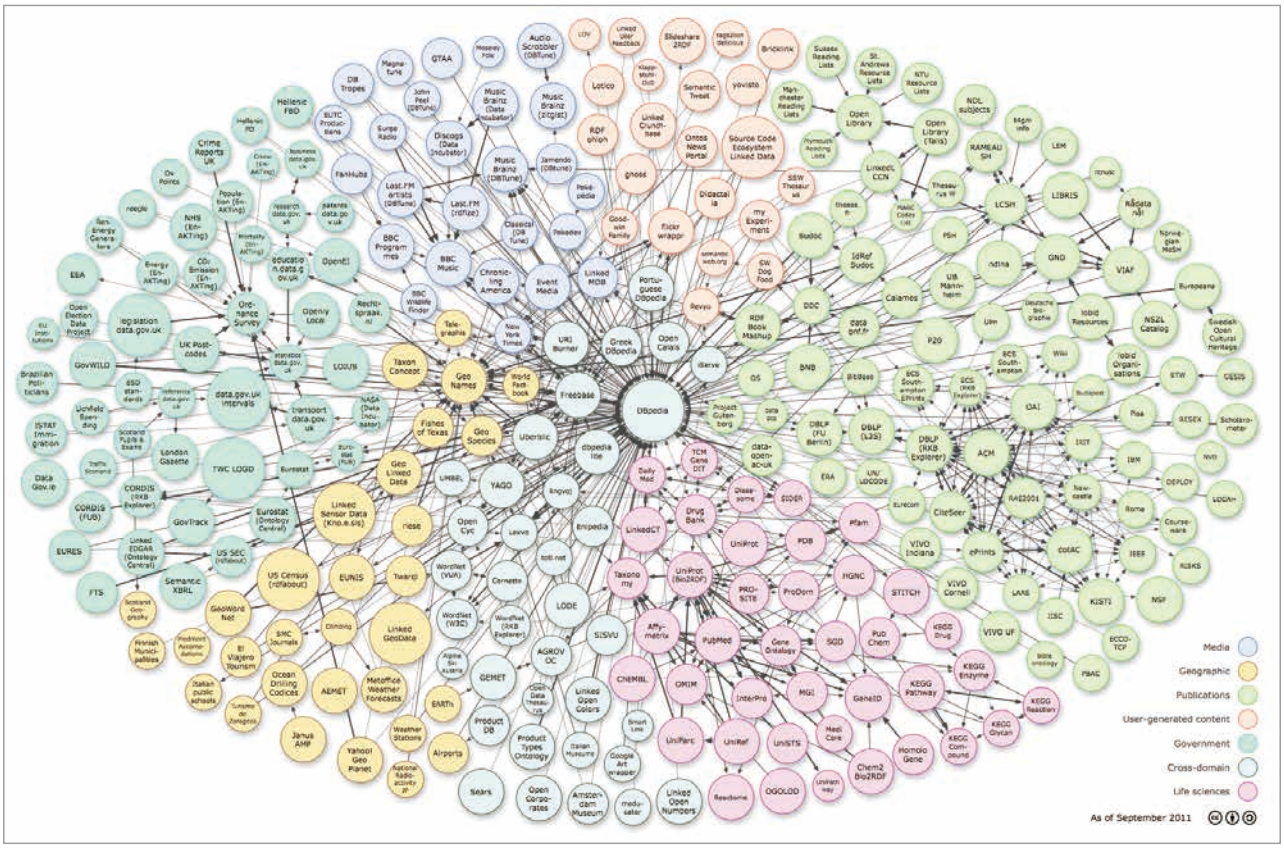

図 -2　LOD cloud diagram 2011 年 9 月版（http://lod-cloud.net）

の中から，所定の条件を満たすものが登録されている．データセットの総数は 2007 年 5 月にはわずか 12 個であったが，2008 年 9 月には 45 個，2010 年 9 月には 203 個，最新版である 2011 年 9 月には 295 個となっており，急速に増加していることが確認できる．また，LOD cloud diagram への掲載を希望するサイトの審査結果が公開されており，未登録のデータセットが 200 以上存在していることが分かる☆5．

データセットの性質や特徴については文献 3）で詳細な分析がなされているので参照されたい．ここでは，データへのアクセス方法に関する調査結果についてのみ述べる．一般に，LOD はファイルまたは SPARQL エンドポイントと呼ばれるデータベースのインタフェースを通じて公開される．SPARQL は W3C によって標準化された RDF データに対するクエリ言語である☆6．SPARQL エンドポイントが利用できる場合にはクエリを記述し，エンドポイントに投入することで必要なデータのみを入手することができるが，提供側があらかじめデータベースを用意する必要がある．ファイルによる公開は提供側にとって容易である一方，データの処理コストを利用者側が負担しなければならない．LOD cloud diagram の分析によれば，295 のサイトのうち 201 のデータセットが SPARQL に対応している．また，Datahub の登録情報の中では 485 の SPARQL エンドポイントが存在している．利用者の求めに応じてデータの提供方法が高度化していることが分かる．

LOD cloud diagram の中心に位置し，多くのリンクを獲得しているデータセットが DBpedia である☆7．DBpedia は Wikipedia のコンテンツに含まれる Infobox に注目し，LOD を自動生成して提供するサービスである．SPARQL エンドポイントも用意されている．Wikipedia 自体が事実情報を収集しており，記事数も多いことから，他のデータセット

☆5 http://validator.lod-cloud.net
☆6 http://www.w3.org/TR/sparql11-overview/
☆7 http://dbpedia.org

が DBpedia の該当エントリにリンクする例が多い．

DBpedia 以外で大規模なデータ公開を行っている例として Europeana がある．Europeana は EU 圏の図書館・博物館・美術館が持つ計 2,000 万以上の作品情報を集約・公開するサービスだが，実験的プロジェクトとしてすべてのデータが LOD 化され，自由に利用することができる☆8．Europeana に限らず，学術分野は LOD cloud diagram の 7 分野の 1 つに数えられるほど積極的な対応がなされており，図書館の総合目録サービス WorldCat や電子ジャーナルの ID を管理する CrossRef などの大規模サービスにおいて書誌情報の LOD 化が進んでいる．

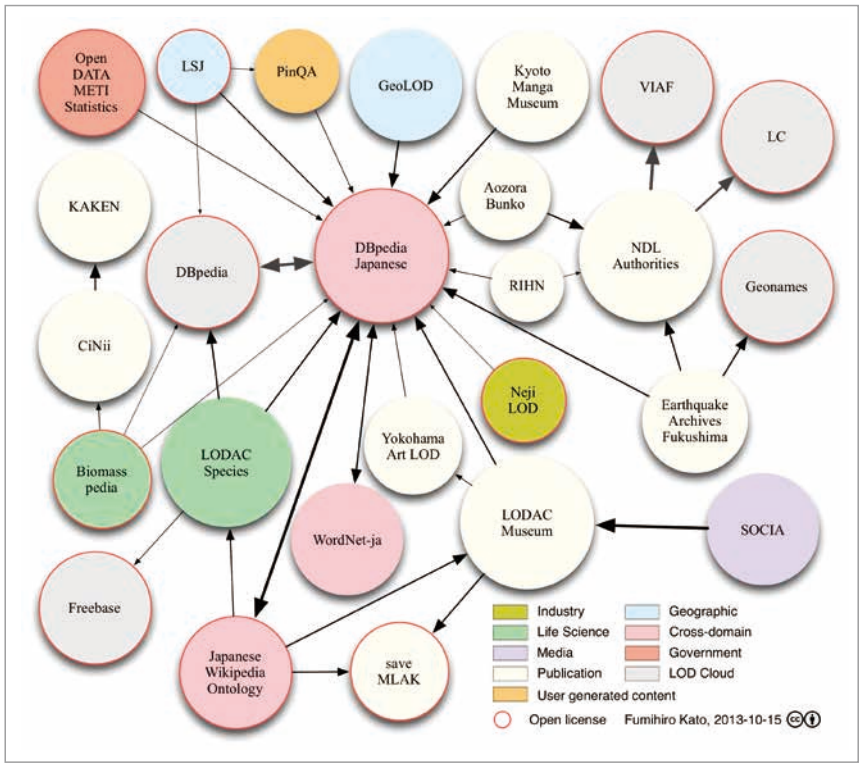

図 -3　日本の LOD 2013 年 10 月版

日本の Linked Open Data

2011 年 9 月版の LOD cloud diagram において，日本からは国立国会図書館による Web NDL Authorities（NDLA：典拠データ検索・提供サービス）☆9 の 1 サイトのみにとどまる．NDLA は蔵書管理のために著者名やキーワードを体系的に整備したもので，これを LOD として利用することができる．SPARQL エンドポイントも提供されており，LOD の先進例の 1 つであるといえる．

世界の潮流と軌を一にして，国内でも Linked Data ないし LOD に対応しているサイトは順調に増加している．

図 -3 は情報・システム研究機構の加藤文彦氏が作成した日本版 LOD cloud diagram である．この図には LOD の定義にあてはまらない，ライセンスが明確でないデータセットも含まれるが，データの Web が着実に育っていることが理解できる．

2012 年 5 月には筆者がかかわる国立情報学研究所 LODAC プロジェクトによって，Wikipedia 日本語版を対象とした DBpedia Japanese がリリースされ，日本の LOD のハブになっている☆10．また，Wikipedia 日本語版のリンク関係から Infobox には現れない概念間の関係を抽出し，利用可能にする Wikipedia オントロジー☆11 や Wikipedia シソーラス☆12，日本語 WordNet の LOD 化などのプロジェクトもあり☆13，基盤となるデータの整備は進んでいるといえる．

データの作成を支援する仕組みとして，理化学研究所ではスプレッドシートをアップロードすると RDF に自動的に変換・公開する LinkData を提供している☆14．LinkData は福井県鯖江市をはじめとする地方自治体の LOD 公開手段としても広く利用されている．

☆8 http://pro.europeana.eu/linked-open-data
☆9 http://id.ndl.go.jp/auth/ndla
☆10 http://ja.dbpedia.org
☆11 http://www.wikipediaontology.org
☆12 http://dev.sigwp.org/WikipediaThesaurusV3/
☆13 http://wordnet.jp/repositories/wordnet-ja
☆14 http://linkdata.org

政府レベルのオープンデータに関しては，経済産業省の Open DATA METI において，試験的に一部のデータを LOD 化し，公開している事例がある☆15．総務省が管轄する統計センターでは，市町村合併などによって複雑な体系を持つ標準地域コードを LOD でモデル化するための検討が行われている．内閣官房が 2013 年度後半に公開するオープンデータカタログでは，メタデータが LOD として利用可能になる予定である．

学術情報分野では大規模なデータベースでの LOD の採用が進んでおり，NDLA に続いて国立国会図書館サーチ・東日本大震災アーカイブ，国立情報学研究所の CiNii・KAKEN などでデータを利用することができる．

Linked Open Data を使う

先に述べたとおり，LOD の技術的なメリットは，標準化された知識表現形式（RDF）とデータアクセス手段（SPARQL）が提供されることにある．これらの標準に基づくライブラリを用いることで，開発者は取得やパースの手間を大幅に軽減できる．また SPARQL についてはオープンソースならびに商用のデータベース（RDF ストア）が多数開発されており，手持ちの LOD を RDF ストアに投入すれば SPARQL エンドポイントとして機能するため，柔軟な問合せができる環境を容易に構築可能である．ここでは LOD を活用したアプリケーションを用途ごとに紹介する．

❖ ブラウズ・検索

LOD はグラフ構造を持つだけでなく，データセット間のリンクも多いことから，全体的なスキーマを把握することは難しい．よってデータの閲覧は探索的にならざるを得ない．Tabulator は LOD の取得と表示に特化したブラウザ用のアドオンであり，LOD に対して Web ページと同じようにクリックによるデータの遷移を可能にする☆16．DashSearch LD は SPARQL エンドポイントに対して対話的に操作を行い，データ構造を理解しながら目的のデータを入手できるよう支援する Web サービスである[4]．また，一般の Web 検索エンジンと同様に，あらかじめ Web 上の RDF や LOD を収集し，キーワードや属性で検索できる Sindice のような検索サービスもある☆17．Sindice には 7 億以上のデータが格納されている．

❖ マッシュアップ

LOD によってデータ形式やアクセス方法が標準化されていることから，既存の情報源を組み合わせたサービス，すなわちマッシュアップの開発コストはきわめて低くなることが期待される．初期のマッシュアップの代表例として，携帯端末向けに地図情報と DBpedia の施設情報を組み合わせた DBpedia Mobile を挙げることができる☆18．DBpedia の情報は SPARQL を通じて入手するため，エンドポイントの設定を変えることで他のデータセットの情報に差し替えることが容易である．

実際に複数のデータセットを利用したサービスを構築した例として，横浜市内の芸術関連情報を地図から探せるサービスであるヨコハマアートスポットがある[5]．国立情報学研究所 LODAC プロジェクト，横浜市芸術文化振興財団，NTT レゾナントがそれぞれ提供している SPARQL エンドポイントから施設・収蔵品・イベント・口コミなどの情報を取得し，ユーザの要求に応じて関連情報を表示する．個々のデータセットは異なる組織によって維持・管理されており，スキーマも大きく異なるが，サービスの実装にあたっては，施設情報の URI を共通化するなどの前処理以外には調整の必要がなく，SPARQL を用いて速やかな開発・提供を行うことができた．

横浜市金沢区のかなざわ育なび .net では，区役所内で部署ごとに管理されているデータを LOD 化し，子育てに必要な情報を部署の枠を超えて一括で検索・閲覧することのできるサービスを提供してい

☆15 http://datameti.go.jp/data/dataset/statistics_kougyou_2010
☆16 http://www.w3.org/2005/ajar/tab
☆17 http://sindice.com
☆18 http://dbpedia.org/DBpediaMobile

る☆19．各々の部署においてデータを作成するためのワークフローやフォーマットを変えることは困難だが，すでに作成されたデータを LOD に変換し，SPARQL エンドポイントに投入する機構を追加することで現場の負荷を最小限にとどめたままデータの標準化を可能にしている．

❖ LOD チャレンジ

LOD の利活用を促進するために，有志によってコンテスト形式の Linked Open Data チャレンジ（LOD チャレンジ）が開催されている☆20．2011 年から開催が始まり，現在（2013 年 10 月）は第 3 回の募集が行われている．LOD チャレンジの特徴は，アプリケーションだけを募集するのではなく，LOD の利活用アイディアやデータを用いた可視化手法も同時に募集することで，開発者でない層に対して門戸を広げていることにある．また，特筆すべきは自作のデータを募集するデータ部門の存在である．一般的に，データ作成はアプリケーション開発と同様に労力が大きいものの評価の対象になりにくい．そこで，LOD チャレンジではデータ公開を活性化させることを目的としてこのような部門が設置されている．

第 1 回はデータ部門が 21 件，アイデア部門が 34 件，アプリケーション部門が 18 件の計 73 件の応募であったのに対して，第 2 回にはデータ部門が 87 件，アイデア部門が 50 件，アプリケーション部門が 44 件，新設の可視化部門が 24 件の計 205 件と 3 倍近い応募があった．応募されたデータを用いて別の参加者がアプリケーションを作成するという，データを介した間接的なコラボレーションが起こるなど，事前の想定を超える成果も生まれている．

データ部門の受賞作品に注目すると，東日本大震災時にボランティアが作成した図書館・博物館などの社会教育施設の全件リストを LOD 化した saveMLAK☆21 や，アニメーション作品の舞台となっている地域の位置情報など☆22，政府・地方自治体にとってデータ作成の負荷がきわめて大きいものや，ポップカルチャーなどの新たな分野のデータが散見されるのが興味深い．

2013 年度は LOD チャレンジ以外にもアーバンデータチャレンジ東京☆23 やオープンデータ・ビッグデータ活用推進協議会のアイデアコンテスト☆24 など多数のコンテストが開催されている．これらは必ずしも LOD に限定されたものではないが，オープンデータの利活用そのものの有効性が評価される時期にあると思われる．

課題と展望

これまで述べてきたように，LOD がもたらす「データの Web」は拡大を続けており，それに伴って利活用の機会も増加している．一方で，LOD はデータセット間にリンクが存在しなければ質的な向上が見込めず，またこの作業が最も困難である．

LOD は原理上すべての事物に URI をつける必要があり，同じ URI を持つリソースはどのデータセットに存在していたとしても同じものであるという唯一名仮説に基づいている．ただし，個々の Web サイトあるいはデータセットがボトムアップに構築される Web において，異なるデータセットに存在する同じ事物に同じ URI がつけられる可能性はきわめて低い．すでに異なる識別子がつけられているが，同じものであると見なしたい一対のリソースがある場合には，それらのリソースを「同じである」または「類似している」という意味のプロパティ，具体的には owl:sameAs，skos:exactMatch，skos:closeMatch などを用いてリンクする必要がある．

しかしながら，何をもって「同じである」と見なすことができるかの判断が難しい．名前あるいはラベルが同一であるだけでは同じものであるとは確定できない場合も多く，このような場合にはデータの内容を逐一確認して同一性の判定を行わなければな

☆19 http://kirakana.city.yokohama.lg.jp
☆20 http://lod.sfc.keio.ac.jp
☆21 http://savemlak.jp
☆22 http://cheese-factory.info/lod.html
☆23 http://aigid.jp/GIS/udct/2013/
☆24 https://www.facebook.com/bigdataopendata4city

らない．いわゆる名寄せ処理を大規模なデータセット群に対して適用するために機械学習などの手法が使われることが多く，一定の成果は得られるものの，少数のエラーが出ることは避けられない．分野によってはエラーが許されないこともあり，その際には人手での確認が必要となるが，コストが増大する恐れがある．一方，名前だけで同一性を判定できるような分野もある．これは対象分野において曖昧性を回避する命名ルールが確立されている場合であり，リンク付けの実行者はその知識を事前に理解しているかどうかによって想定される作業の規模が大きく変わる．

同一性の問題以外にも課題は多い．LOD はグラフ構造を持つが，現在オープンデータの一環として提供されるデータは統計表などスプレッドシートで作成されたものが多い．統計表のデータをシリアライズするには SDMX と呼ばれる規格があり☆25，このモデルを LOD で扱うための Data Cube 語彙が提案されている☆26．しかし，一般的に表形式のデータをグラフに変換すると表現が冗長になり，可読性が下がる傾向にある．そのため，このような表形式データの LOD 化に際しては，そのメリットを理解し，必要に応じてメタデータだけを LOD 化するなど，コストパフォーマンスを考慮した対応が必要になる．逆に，スプレッドシートのソフトウェアはその柔軟性の高さゆえに曖昧な構造のデータを作ることも可能である．その際には本来の構造を表現できるようなモデル化を行わなければならず，元のデータから自動的に変換が可能かといったことを検証する必要がある．

これらの問題に対しては，分野に関する専門知識を持ち，かつコンピュータに精通した人物が適切に判断し，開発者と分業しながら大規模かつ継続的にデータを維持管理する体制が作られることが望ましい．そのためには，「データキュレーション」とも呼ぶべきスキルセットを定義し，そのようなスキルを持つ人材を組織的に育成する必要がある．LOD のポテンシャルを発揮させるためには，このコストを誰がどのように負担すべきかが大きな課題となる．LOD のもとになるデータがオープンであれば，必ずしもデータキュレーションはデータの持ち主の仕事ではなく，ビジネスとしてアウトソースされるものになる可能性もある．

近年，Web 上の構造化データは質・量ともに向上しており，Web 検索エンジンにおいては Google の Knowledge Graph に代表されるセマンティクスに基づくナビゲーションの提供事例や，IBM の Watson のように特定分野の Q&A を人工知能技術で行うための基礎データとして取り入れられることも増えている[6]．LOD ならではのキラーアプリケーションが求められている状況ではあるが，地道かつ着実なデータ整備の活動との両輪で進めていくことが必要であろう．

☆25 http://sdmx.org
☆26 http://www.w3.org/TR/vocab-data-cube/

参考文献

1) Bizer, C., Heath, T. and Berners-Lee, T. : 萩野達也（翻訳）: Linked Data の仕組み，情報処理，Vol.52, No.3, pp.284-292 (2011).
2) Berners-Lee, T. : Linked Data - Design Issues, http://www.w3.org/DesignIssues/LinkedData.html (2006).
3) Bizer, C., Jentzsch, A. and Cyganiak, R. : State of the LOD Cloud, http://lod-cloud.net/state/ (2011).
4) Goto, T., Takeda, H. and Hamasaki, M. : DashSearch LD : Exploratory Search for Linked Data, Proceedings of the 2nd Joint International Semantic Technology Conference (2012).
5) 松村冬子，小林巌生，嘉村哲郎，加藤文彦，高橋　徹，上田洋，大向一輝，武田英明 : Linked Open Data による博物館情報および地域情報の連携活用，情報処理学会人文科学とコンピュータシンポジウム論文集，pp.403-408 (2011).
6) Ni, Y., Zhang, L., Qiu, Z. and Wang, C. : Enhancing the Open-Domain Classification of Named Entity Using Linked Open Data, The Semantic Web - ISWC 2010, pp.566-581 (2010).

（2013 年 10 月 14 日受付）

■ **大向 一輝**（正会員）　i2k@nii.ac.jp

1977 年京都生まれ．2005 年総合研究大学院大学博士課程修了．博士（情報学）．2005 年国立情報学研究所助手，2009 年同准教授．セマンティック Web やソーシャルメディア，オープンデータの研究開発に携わる．人工知能学会会員．