拡大するLinked Open Data

武田 英明 ●国立情報学研究所 教授

クロスドメイン関係や、図書館関係、バイオサイエンス関係、政府データ関係のアーリーアダプター事例が先導。日本の政府標準利用規約もCC-BY 4.0互換に。

■データのWeb

Webは1990年代初頭に学術情報の共有のために発明されたが、瞬く間に学術世界のみならず社会全体に普及し、われわれの情報流通の仕組みを大きく変化させた。このWebによる情報流通の革命と同じことが今、データの流通に起きつつある。それがLinked DataあるいはLinked Open Data（LOD）である。

これまでのWebはおもに文書的情報を相互にリンクしてネットワークを作っている。いわば「文書のWeb」といえる。これに対して、LODは同様のネットワークをデータの間で作る。このため、LODは「データのWeb」と呼ばれる。ただし、「文書のWeb」ではリンクは種類のないものであったのに対して、「データのWeb」ではデータ間の関係を示すラベルをもつリンクになる。これにより、「このデータとこのデータは同一」「この項目の定義はこれ」といった関係を作ることができる。

LODはWebと同じようにグローバルに共有するデータ空間である。すなわち、データが世界のどこのだれが管理しているデータセット（データベース）に含まれているかを意識することなく、アクセスしたり、リンクしたりできる。これまであったデータセットとデータセットの間の障壁は存在しない。個々のデータセットはお互いにリンクし合うことで、一つのグローバルなデータセットの一部となるわけである。

■Linked Open Data（LOD）の仕組み

じつはLOD技術は、セマンティックWeb研究で作られた技術をデータの表現に利用したもので、多くの標準がすでに整備されている。セマンティックWebとはTim Berners-Lee氏が提唱した、現行のWebより高度に知識を記述できるWebを作るというビジョンである。そのポイントはWebのグローバルな情報共有空間はそのままに、その上に標準的なメタデータの記法であるRDF（Resource Description Framework）やそのスキーマを記述する言語（RDF SchemaやOWL）を用意することで、グローバルに知識を共有する仕組みである。

まず、RDF（Resource Description Framework）という言語ですべてを書く。RDFは「主語」「述語」「目的語」に相当する3語の組み合わせ（3つ組）ですべての情報を書く言語である。たとえば「AさんはBさんを知っている」という関係は「A knows B」という3つ組で表現される。このとき表現したい事物（たとえば、前述の例ではA、B）には個別のURIを与える。ここでは個々の事物に

インターネット白書／ ©1996-2016 Impress R&D

URIを振ることで、URIが世界中でユニークなIDとして使えることが重要である。

さらにデータを記述する様式が決まっている場合、その様式をスキーマとして別途定義する。RDFを拡張したRDF Schema（RDFS）という言語では、データ記述の様式をクラスとプロパティの組み合わせで定義する。クラスはある事物や事象を示し、プロパティはクラスがもつ属性やクラス間の関係の型を示す。RDFSはRDFの拡張であるので、クラスやプロパティの定義もURIを与えて、参照可能にする。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ない。すでに他で定義されたスキーマが使えるあるときは、それを使うことができる。

URIとRDFを使うことでデータはデータがどこに含まれているといったことを気にせずに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4つの原則を提唱している。

1. 事物をURIを使って名前付けしよう
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

この原則に基づくとデータセットは相互につながり、データセットのネットワークができる。

■LODが作る世界

現在、上記の原則に基づいたLODが多数公開されている。資料4-2-1に示すのは、世界中のLODのネットワークを示す図で、これをLOD Cloudと呼んでいる。2014年8月の時点での図で、500以上のデータセットが含まれている。丸が個別のデータセットを示し、データセットとデータセットを結んでいる線は、データセット内のデータ同士にリンクがあることを示している。

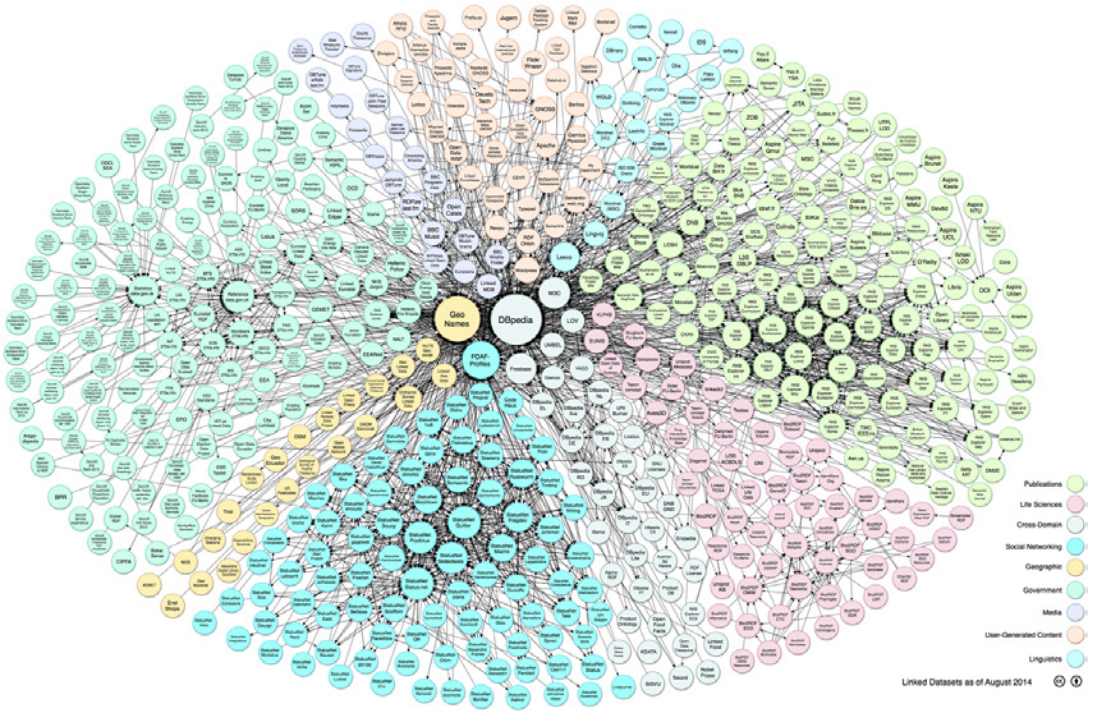
中心にあるのはDBpediaというもので、Wikipediaの情報をLOD化したデータセットである。Wikipediaは百科事典なので、さまざまな分野の項目が含まれている。そのため、他の多くのデータセットと結びつきやすく、この図の中心にある。

資料4-2-1ではドメインごとにデータセットが色分けされている。右上の部分は出版物に関連するデータセットで、図書館の書籍書誌データやジャーナル論文の書誌情報といったものが含まれている。右下はバイオサイエンスなどを含むライフサイエンスに関係するデータセットである。左端は政府関係のデータセットである。ここには、アメリカの政府データポータルであるdata.govやイギリスのポータルdata.gov.ukのデータのうち、LODのものが含まれている。

これらの異なるドメインのデータがDBpediaを中心にゆるくつながっていることがわかる。これらはみな前述のRDFで書かれているので、異なるデータセットのデータであってもあわせて使うことができる。

インターネット白書／ ©1996-2016 Impress R&D

資料4-2-1　世界中のLODのネットワークを示す「LOD Cloud」



出典　Linking Open Data cloud diagram 2014[1]

■LODの使い方

LODはこれまでのWebと同様に、検索サービスを通じてアクセスして使うこともできるし、リンクを順に辿ってみるというブラウジングでアクセスして使うこともできる。

しかし、LODの特徴を活かした使い方は、マッシュアップアプリケーションを通じての使い方である。通常のマッシュアップアプリケーションでは、複数のWebサービスのWeb APIを利用してアプリケーションを構築する。LODを使ったマッシュアップアプリケーションは、通常のマッシュアップアプリケーションに比べ、より容易に構築できる。

LODではデータのアクセスの仕方が統一されているため、Web APIの場合のように個別に対応を変える必要がない。また、記述形式もRDFで統一されているので、アプリケーションの中でも統一的に扱うことができる。さらに異なるデータセット間でもリンクがあれば、そのまま使うことができるため、データ統合の手間を省くことができる。

■LOD技術の展開

LOD技術の利用はデータの公開の場面に限定されているわけではない。データモデルをRDFのみとすることで、多種多様なデータを統合的に処理するETLの仕組みとして使うこともできる。データの収集、洗練、統合、分析、公開、利用といった一連のプロセスをRDF処理だけで実現できる。その際、データやスキーマはそのまま外部から取り込むことができ、処理結果のほうもそのまま外部へ提供することもできる。また、さまざまなRDFに関わるツール、ソフトウエアをその間で使うことができる。

インターネット白書／ ©1996-2016 Impress R&D

■世界での動向

LODに関する技術としては、2012年から2014年までに一通りの技術標準（RDF、Turtle、JSON-LD、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やアプリケーション、サービスが作られ、アーリーアダプター段階にあるといえる。一方、これらのイノベーターやアーリーアダプターの経験から、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。

新たな技術要素としては、Linked Data Platformや、CSV on the Web、RDF Data Shapesなどがある。これらはみな、実際のアプリケーションやサービスを作るときに不足する機能として提案されたものである。

Linked Data Platformは、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。CSV on the Webは、実際のデータで多く見られるテーブル形式のデータをLODで扱いやすくするための標準で、2015年12月W3C勧告となっている。このCSV on the Webでは、CSVファイルの記法やメタデータ語彙、JSON変換方法、RDF変換方法など規定されている。RDF Data Shapesは、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記述する規定するもので、現在はまだ議論中である。

アーリーアダプターとしての適用事例は、クロスドメイン関係や、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している。

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。クロスドメイン関係では、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）。

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と同様に共同作業で構築していこうというWikidataプロジェクトが進展している。これもLODになっている。

図書館における書誌および典拠は元々公開・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。そのため、LODと相性がよく、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（LCSH）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種類の情報がLODとして公開されている。欧州では英国図書館（2011年7月）、ドイツ国立図書館（2012年1月）、スイス国立図書館（2012年9月）、フランス国立図書館（2014年1月）、スペイン国立図書館（2014年11月）によるデータ公開が行われている。

博物館関係でも同様に公開が進んでいる。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。ヨーロッパの図書館や博物館等のデータを収集・公開するEuropeanaでも、ライセンスがオープンになったものを順次LODとして公開している。またゲッティ研究所は、2014年から2015年にかけて美術関係のシソーラス（AAT）と地名典拠（TGN）、芸術家典拠（ULAN）を順次LOD化して公開した。

バイオサイエンス分野では、元々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）が利用されている。そのため、LODとの親和性が高い。このことから、米国国立医学図書館（NLM）が公開する生命科学用語集MeSHをはじめとして、多くのデータがLOD化されている。バイオサイエンス分野では多数のデータベースがあり、それらを横断的につなぐの

インターネット白書／ ©1996-2016 Impress R&D

にLODは有効である。たとえばbio2rdfプロジェクトでは、19のデータセットをRDF化してつないでいる。

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（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。このDCLGやグレーター・マンチェスター、スコットランド政府等は、研究プロジェクトの成果を用いて構築されたSwirrl社のLinked Data公開プラットフォームPublishMyDataを用いて、統計データなどをLODとして公開している。

■国内での動向

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。国内でもアーリーアダプターとしてデータ公開やサービス、アプリが多く出現している。

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は早くから簡便なLOD化ツールを提供してきた。またjig.jp社は、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。この他、Datashelf（インフォランジ社）などもある。

国内ではクロスドメイン関係と図書館関係、政府データ関係が中心に適用が進んでいる。

クロスドメイン関係では、2012年より、日本語WikipediaをLOD化したDBpedia Japaneseは公開されていた。2015年5月からは、地名のURI基盤としてgeonames.jpがスタートしている。また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およびXMLで提供している。

図書館関係ではすでに国会図書館の典拠データ、国立情報学研究所の論文検索サービスCiNiiのRDF化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化もスタートしている。

政府データ関係では地方自治体のオープンデータのLODとしての公開が増えてきている。多くの自治体はLinkData.orgのような既存のプラットフォームを介して公開しており、北海道森町のように共通語彙基盤の応用をLinkData.org上で試みる例も出てきた。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。また関西では、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。さらに、これらのデータを使ったアイデアソンやハッカソンも多く開催されている。

なお、日本政府は、2014年に制定した政府標準利用規約（第1.0版）を2015年12月に第2.0版として改定し、CC-BY 4.0互換とした。このため、今後、地方自治体を含む政府関係のより多くのデータがLODとして公開されることが期待される。

インターネット白書／ ©1996-2016 Impress R&D

1. Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. http://lod-cloud.net/

インターネット白書／ ©1996-2016 Impress R&D

インターネット白書 ARCHIVES

1996,1997,1998,1999,2000....

［インターネット白書ARCHIVES］ご利用上の注意

このファイルは、株式会社インプレスR&Dが1996年～2016年までに発行したインターネットの年鑑『インターネット白書』の誌面をPDF化し、「インターネット白書ARCHIVES」として以下のウェブサイトで公開しているものです。

http://IWParchives.jp/

このファイルをご利用いただくにあたり、下記の注意事項を必ずお読みください。

- 記載されている内容（技術解説、データ、URL、名称など）は発行当時のものです。
- 収録されている内容は著作権法上の保護を受けています。著作権はそれぞれの記事の著作者（執筆者、写真・図の作成者、編集部など）が保持しています。
- 著作者から許諾が得られなかった著作物は掲載されていない場合があります。
- このファイルの内容を改変したり、商用目的として再利用したりすることはできません。あくまで個人や企業の非商用利用での閲覧、複製、送信に限られます。
- 収録されている内容を何らかの媒体に引用としてご利用される際は、出典として媒体名および年号、該当ページ番号、発行元（株式会社インプレスR&D）などの情報をご明記ください。
- オリジナルの発行時点では、株式会社インプレスR&D（初期は株式会社インプレス）と著作権者は内容が正確なものであるように最大限に努めましたが、すべての情報が完全に正確であることは保証できません。このファイルの内容に起因する直接的および間接的な損害に対して、一切の責任を負いません。お客様個人の責任においてご利用ください。

お問い合わせ先　株式会社インプレスR&D　✉ iwp-info@impress.co.jp

©1996-2016 Impress R&D, All rights reserved.