

知っておきたいキーワード

LOD (Linked Open Data)

(正会員) 宮崎 勝[†]

[†] NHK 放送技術研究所

"Linked Open Data" by Masaru Miyazaki (NHK Science & Technology Research Laboratories, Tokyo)

キーワード：LOD, RDF, リンクトデータ, オープンデータ, セマンティック Web

まえがき

インターネット上では多くのページがハイパーリンクを介して蜘蛛の巣のようにつながり、巨大な情報空間である Web を形成しています。Web は長

年にわたり「人間が情報にアクセスするため」に発展してきました。この Web を単なる Web ページの集合体ではなく、オープンライセンスのデータを含んだグローバルなデータ空間として捉え、そのデータの「ソフトウェア

による利活用」を実現する技術が LOD (Linked Open Data) です。本稿では、LOD の基本的な考え方について解説します。

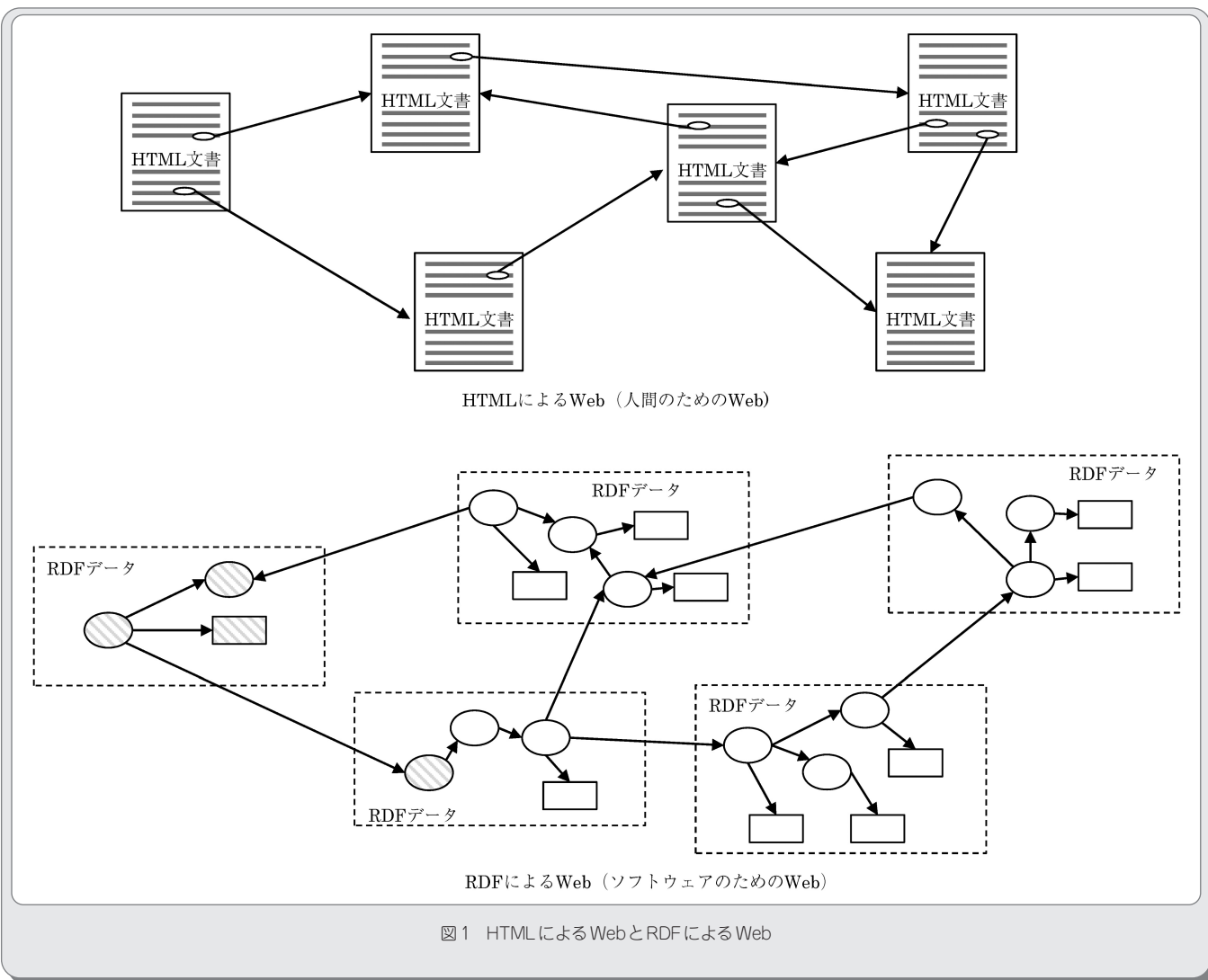
データの Web

1990 年代に登場した World Wide Web により、誰もが簡単に Web 上のさまざまな情報を蓄えた空間にアクセスできるようになりました。2000 年以降、Web に膨大な情報があふれるようになると、Web を巨大な知識のデータベースとして捉え、その情報を人間だけではなく、ソフトウェアが自動処理できるようにするという構想が生まれました。この考え方は「セマンティック Web」と呼ばれ、AI (Artificial Intelligence: 人工知能) 関連技術を背景として、自動処理可能なデータ記述手法などの規格化・検討が進められています。

Web ページを記述する HTML (Hyper Text Markup Language) では、

ページのレイアウトなどの指定とともに、ハイパーテキストにより他の文書へのリンク情報を記述することができるため、これをたどることで、ユーザは情報空間の中から目的の情報を見つけることができます。しかし HTML では、文書中の個々のオブジェクトの意味や、他の文書へのリンクの意味は明示的に示されていません。例えば「バラク・オバマ」という文字列が Web ページ上にあった場合、人間はそれを見て「あの」アメリカの大統領のことだと容易に想像し、Web ページ上にはその人物に関する情報が記載されていることをすぐに理解するでしょう。しかし、この理解をソフトウェア処理で実現するためには、この文字列が地名や組織名ではなく「人名」を表しており、さらにその人物の「職業」が「大統

領」であるということを、ソフトウェアが解釈できる形で書いておかねばなりません。このように、ソフトウェアが Web の情報空間をたどって求める情報を探し、利用するためには、データの意味的情報を明確に記述し、さらに別のデータとの関係性を、明確な意味を持ったリンク情報で表現する技術が必要となります。これが Linked Data です。Linked Data では、RDF (Resource Description Framework) という意味のあるリンク構造を表現できる形式でデータを記述します¹⁾。これにより、人間が利用するための HTML による Web とは異なり、ソフトウェア処理可能なデータの Web 空間を記述することができるのです (図 1)。



RDF

Linked Dataは、前述したRDFで記述されます。RDFではWeb上にあるすべてのデータを「リソース」と呼び、そのリソースに関する情報を記述するための枠組みを提供します。リソースの情報は、「主語 (Subject)」、「述語 (Property)」、「目的語 (Object)」という3つの要素からなる「トリプル」と呼ばれる形式で記述されます。また、すべてのリソースは必ずURI (Uniform Resource Identifier : Web上で情報を一意に指し示す住所のようなもの) で記述されます。これにより、ソフトウェアはリソースの情報を一意に特定し、そのリソースに関する情報を取得することができます。

図2は、xyz社が、社員である田中太郎さんに関する情報をRDFトリプルで記述する場合の例を示しています。

図2 (a) は、田中太郎さんの年齢が25歳であるということを記述するトリプルです。主語になっているのは、情報の記述対象となっている「田中太郎」を世界中で一意に指し示すURIです。これを単に「田中太郎」という文字列で記述してしまうと、一体どこの田中太郎さんについての記述なのかわからなくなってしまうので、URIを用いて一意に指定します。ここでは、xyz社が独自に定めた“http://www...xyz.com/person/taro2435”というURIを用いています。

述語もURIで記述されます。述語に関しては、年齢を表すものをxyz社で

独自に決めてもよいですが、ここではあえて、FOAF (Friend of a Friend) というプロジェクト²⁾で定義されたageという名前の述語をURIで指定しています。FOAFで定義された述語は人物の情報を記述するために広く一般的に使われているため、ソフトウェアは「人物の年齢」を示す述語であることを正しく理解することができます。このように、広く知られている共通の述語を積極的に利用することで、独自に述語を定義したときよりも、より多くのソフトウェアがその意味を理解して処理を行うことができます。

このトリプルの目的語は、「25」という数値になっています。このように数字や文字列といった具体的な値を表す形式は「リテラル」とい

☞ 呼ばれます。

さて、RDFを用いて主語、述語、目的語という3つ組のデータを作ること、**「xyz社の田中太郎の年齢は25歳である」**という情報を記述することができました。ここで例えば、田中太郎さんの情報として新たに**「所属」**という情報を追加したい場合を考えます。この場合も、図2 (b) のような新たなトリプルを1つ追加するだけです。図2 (b) では、目的語として数値や文字列といったリテラルではなく、**「開発部」**を表すリソースを指定しています。このように、RDFでは、目的語としてリソースを取ること

もできます。広く一般的に使われているRDB (Relational Database) でデータを管理している場合、このように新たな項目を追加するためには、データ構造を変更して、新たに**「所属グループ」**のフィールドを追加する必要があります。RDBなどのデータベースシステムと比べて、RDFは柔軟に情報の追加が可能なおわかりいただけるかと思います。

図2 (a) および図2 (b) では、xyz社の社内情報同士のリンクづけを行っていましたが、RDFでは、目的語として外部にあるリソースを指定することもできます。つまり、自前のデータを、

他者のデータに結びつけることも容易にできるのです。図2 (c) のトリプルは、**「田中太郎さんの居住地が、横浜市の青葉区である」**ことを示していますが、目的語としてWikipedia日本語版の情報をLODとして公開している**「DBpedia Japanese」**³⁾のリソースを指定しています。これにより、このトリプルを利用するソフトウェアは、DBpediaの情報を元に、田中太郎さんの居住地に関する情報を正確に取得し、さまざまな処理を行うことができるようになります。

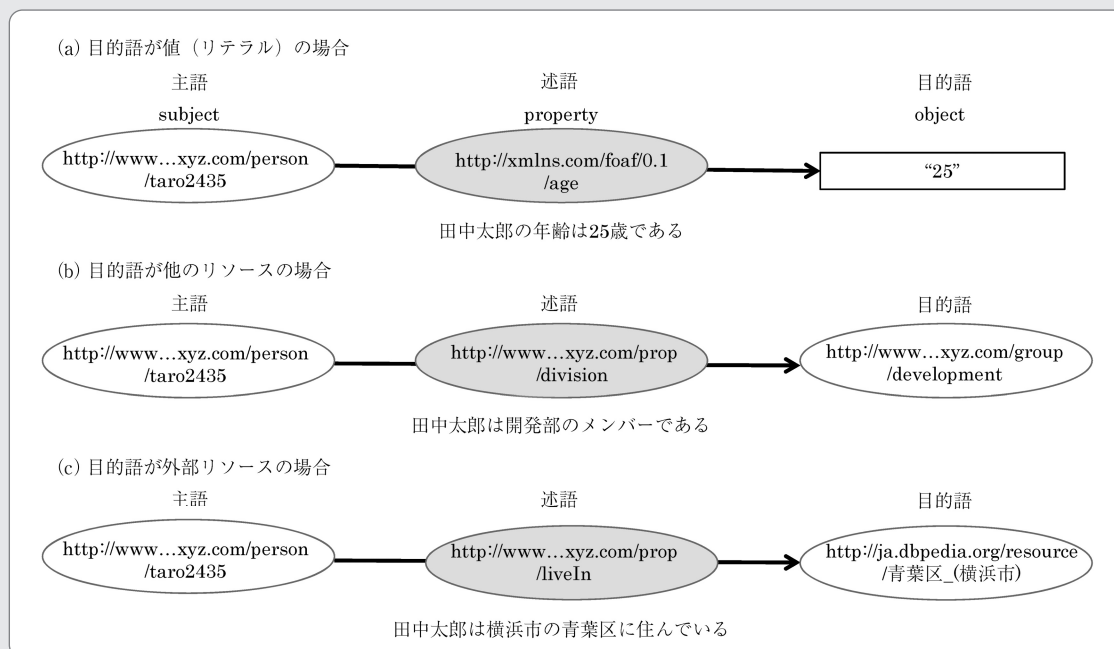


図2 RDFの基本構造 (トリプル)

Linked Dataの4原則

このように、Linked Dataの考え方では、あらゆるデータを意味のある関係でつなぐことができる枠組みが提供されています。WWWやセマンティックWebの考案者であるTim Berners-Leeは、前述したようなLinked Data構築の手順を、図3のように4つの原則として定義しています⁴⁾。

- (1) あらゆるデータの識別子としてURIを使用する。
- (2) 識別子にはHTTP URIを使用し、参照やアクセスを可能にする。
- (3) URIにアクセスされた際には、有用な情報を標準的なフォーマット (RDF など) で提供する。
- (4) データには他の情報源における関連情報へのリンクを含め、Web上の情報発見を支援する。

図3 Linked Dataの4原則

(1) はあらゆるデータ (リソース) を、世の中で一意に識別することが大事である、という考え方を示しています。

また、(2) は、そのリソースに関する情報を取得する際には、よく知られたHTTPプロトコルを使った ☞

☞ 単純かつ汎用的な方法を提供すべき、というものです。(3)は、RDFトリプルのような、データを構造化する柔軟かつ規格化されたデータモデルを利用することを、(4)は、自前の情報

だけでなく、他者の情報へのリンクを積極的に張ることで、データのWeb空間構築を促進する考え方を示しているといえるでしょう。
これら4つの原則にしたがってデー

タを作っていくことで、ソフトウェアがWeb上にあるさまざまなデータを組合せて処理できる世界を実現することが可能となります。

LODとは？

ここまでで、Linked Dataの基本的な考え方について解説しました。では、Linked Open Dataとはどのようなものなのでしょうか？

LODは、Linked Dataの中で、誰でも利用できるようなオープンライセンスになっているもののことを指します。図4は「LODクラウド」と呼ばれる、現在LODとして提供されている主なデータがどのような規模、関係性を持っているかを示す概念図です。2014年8月時点では570個のデータが示されており、さまざまなデータが、分野の壁を超えてつながっている現状を示しています。

近年各国では、政府が持つデータを積極的にオープン化していこうという考え方が活発になっています。2013年のG8サミットで合意された「オープンデータ憲章」でも、「オープン

データはこの世界的な動向の中心に位置している」として明記されています。このような状況に後押しされ、

Linked Open Dataの分野は、ますます発展を続けています。

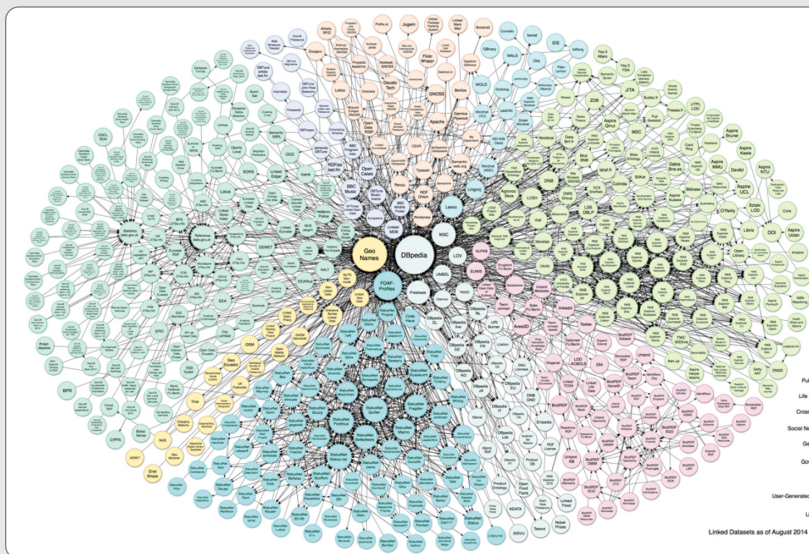


図4 LODクラウド

(出典: "Linking Open Data cloud diagram 2014, by Max Schmachtenberg, Christian Bizer, Anja Jentzsch and Richard Cyganiak. <http://lod-cloud.net/>)

むすび

本稿では、LOD(Linked Open Data)について簡単に解説しました。日本でも、「LODチャレンジJapan」⁵⁾などのコンテスト型のイベントを通じて、さまざまなLODやそれを利用したアプリケーションなどが生まれています。今後、データのWeb化がより進み、LODを基盤としたさまざまなサービスの創出が期待されます。
(2015年10月20日受付)

参考文献

- 1) 神崎正英: “セマンティック・ウェブのためのRDF/OWL入門”, 森北出版(2005)
- 2) <http://www.foaf-project.org/>
- 3) <http://ja.dbpedia.org/>
- 4) Christian Bizer, Tom Heath, Tim Berners-Lee, (翻訳) 萩野達也: “リンクするデータ(Linked Data)～広がり始めたデータのクラウド～: 1. Linked Dataの仕組み Linked Data-The Story So Far”, 情報処理, 52, 3, pp.284-292 (Mar. 2011)
- 5) 乙守信行, 中辻真, 萩野達也: “オープンデータの普及促進を加速させるコンテストの開催～LODチャレンジJapanの取組み～”, 人工知能, 30, 5, pp.598-604 (Sep. 2015)



みやざき 勝

1997年、東京工業大学大学院総合理工学研究科知能科学専攻修了。同年、NHK入局。名古屋放送局を経て、2000年より、放送技術研究所に勤務。

2013年、東京工業大学大学院イノベーションマネジメント研究科技術経営専攻修了。2015年、東京工業大学大学院総合理工学研究科知能システム科学専攻博士課程修了。現在、放送技術研究所ハイブリッド放送システム研究部にて、知識処理を用いたコンテンツ活用技術、放送通信連携サービスの研究に従事。技術経営修士(専門職)、博士(工学)、正会員。