CA1746 [redacted]

動向レビュー

Linked Data の動向

1. はじめに

Linked Data はデータの共有の新しい方法として近年認知されつつある。特にデータのオープン化（オープンデータ）の標準的方法として使われるようになっている。図書館の世界においても所蔵データや件名標目表を Linked Data として公開する図書館が相次いでいる。

本稿では Linked Data の基本的な考え方と全体的な動向・傾向について述べる(1)。

2. Linked Data とはなにか

Linked Data とは、一言で言ってしまえば、データ版の World Wide Web（WWW、以下 Web）である。現在の普通の Web の主たる対象は、人間が理解する文章、文書であり、それがハイパーリンクでつながっているので、「文書の Web」（Web of Documents）といえる。Linked Data は文書ではなくデータがハイパーリンクでつながったもので、「データの Web」（Web of Data）というわけである。

Web が HTML という標準言語を必要としたようにこの「データの Web」にも標準言語が必要であり、それが RDF（Resource Description Framework）である。Linked Data とは、様々な情報源のデータが RDF で記述され、それらが結びついてつくられるデータの集合である。

RDF は元々はメタデータ記述言語であるが、Linked Data ではこれを使ってデータを記述する。RDF では、データは（主語、述語、目的語）という単純な関係として記述される。この一組のデータを RDF 文（RDF Statement）あるいはトリプルと呼ぶ。

パターン化されているデータは RDF スキーマ（RDF Schema：RDFS）を使って、データ構造を明示的に定義して、個別のデータはスキーマ（あるいはクラス）のインスタンスとして記述される。RDF を使うことでデータを一つの標準言語で記述することができる。

しかし、これだけでは単にデータをある言語で記述しただけに過ぎない。Linked Data ではその名の通り、“Link” されないといけない。そこで重要になってくるのが URI（Uniform Resource Identifier）である。URI は URL の拡張として提示されるもので、Web 空間でリソース（資源）を一意に指定することのできる識別子である。URL も URI としてみることができるが、URL と異なり URI はそこに何か（URL でいえば Web 文書）があることを保証するわけではなく、あくまで一意に指し示す識別子である。

RDF ではその主語は URI である必要がある。また述語、目的語も URI でよい。すなわち、RDF を使ってデータを記述する場合、常に Web 空間で一意に識別可能な形で書くということである。さらに目的語として任意の URI が使えるので、自分のデータセットの中の項目を指し示すだけではなく、他のデータセットの中の項目も指し示すことができる。この仕組みによってデータセットを超えて相互に参照しあう Linked Data が可能になる。

Web の創始者であるバーナーズ・リー（Tim Berners-Lee）は Linked Data の 4 原則として以下のものを挙げている(2)。

① ものの名前として URI を使うこと
② ものの名前を調べられるように HTTP URI を使うこと
③ URI を見に行ったとき、RDF や SPARQL のように標準技術によってそれに対する有用な情報を提供できるようにすること
④ より多くのものが発見できるように、データの中に他の URI へのリンクをいれること

何らかのものを言及するときはそれに URI を用意しましょうということである。これにより Web 上で一意にそのものを指し示すことができるようになる（①）。

さらに URI の中でも HTTP URI を使うことで、通常の Web と同じような方法でデータにアクセスできるようになる（②）。

URI というのは識別子に過ぎず、その URI にアクセスするとデータ自身が手に入るようにしておく必要がある。その一つの方法は通常の Web が HTML 文書を返すに対して、Linked Data の URI は RDF 文を返す方法である。あるいは RDF データベースに対する問い合わせ言語 SPARQL（リレーショナルデータベースに対する SQL のようなもの）を使って、問い合わせができるようにしておいてもよい（③）。

そして、そのデータもそのサイト内のデータのみ参照するのではなく、外部のサイトのデータも参照するようにすべきである（④）。

3. LOD クラウドの現状

LOD とは Linking Open Data または Linked Open Data(3)のことを指す。前者であればオープンなデータのつながりを指し、後者であればオープンに利用可能な Linked Data を指すが、あまり指すところの差はない。その Linked Data 間の相互関係を図示したものが LOD クラウドである。

LOD クラウドとはデータサイトの作るネットワー

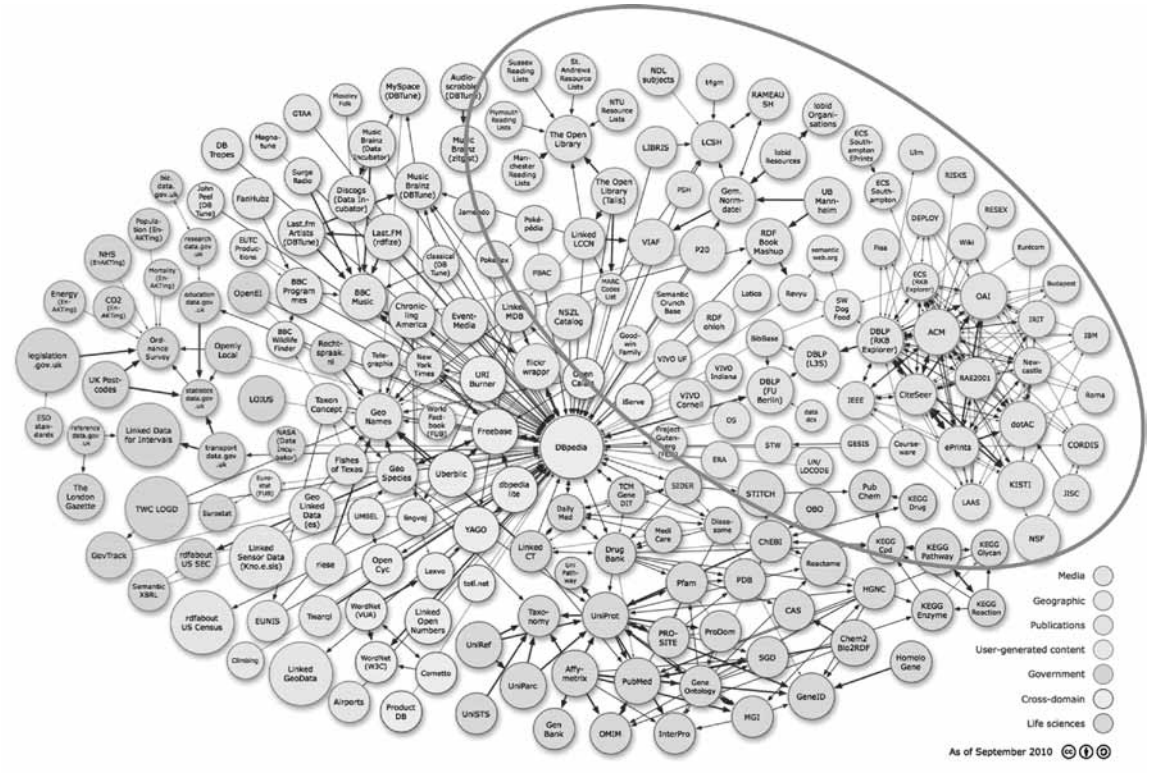

図1　LODクラウド

出典：(5)を基に筆者が加筆

ク図である。Linked Dataの原則のところで述べたように､Linked Dataの強みは異なるデータサイトのデータがつながりあうことができる点である。LODクラウドはその広がりを視覚的に表現したものである。2010年9月時点での状況を図示したものが図1である[(4)]。一つ一つのノードがデータセットを示し、ノード間のリンクはそれらのノード間にデータの参照があることを示す。

中心にDBpedia（WikipediaをLOD化したもの）がある。右上を中心に図のノード全体の1/4以上を占めているのが出版・論文・図書館関係（publication）である（図中で楕円で囲んだ部分の大部分)。ここから反時計回りにみていくと、左上の1/8程度の部分がメディア関係である。左端あたりにあるのが政府関係データ､左下に地理関係とクロスドメインが順にある。右下にありノード全体の1/4弱を占めるのが生命科学関係である。

以下ではバイザー（Chris Bizer）らの分析を中心にして、このLODに含まれるデータが何であるかをみていく[(6)(7)]。

2010年10月時点で全体で約286億トリプル、207データセットである[(8)]。量の割合でみると、政府関係がもっとも多く全体の約41%を占める。ただし、政府関係はカテゴリとしては2010年に初めてできたものである。次は地理関係で約21%、以下はクロスドメイン、生命科学、メディア関係、出版・論文・図書館関係の順に続く。

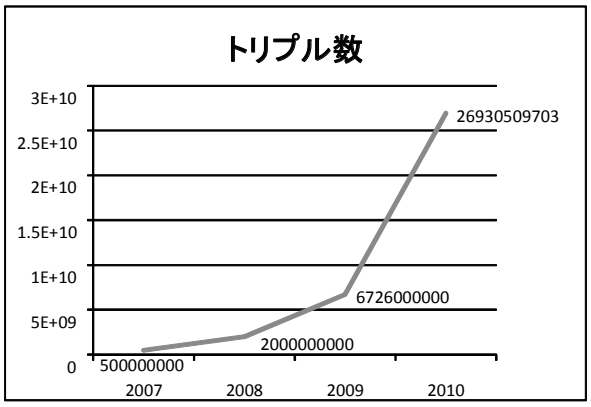

図2　LODのデータ量の変遷

出典：(9)を基に筆者が作成

増加率でみると、2007年から2010年にかけて毎年おおよそ300%ずつ増加している[(10)]。すなわち指数的に増加している（図2)。2009年6月から2010年11月の変化を分野別にみると、出版・論文・図書館関係はおおよそ1,000%の増加をして22億トリプルである[(11)]。2010年に初出の政府関係をのぞけば最大の伸び率である。2010年にこの分野でLODが多くの注目を集め、実際にデータがでてきたことを示している。

出版・論文・図書館関係の分野に関しては、

・米国議会図書館（subject headings）
・ドイツ国立図書館（PND dataset and subject headings）
・スウェーデン国立図書館（Libris-catalog）
・ハンガリー国立図書館（OPAC and Digital Library）
・ドイツ経済学中央図書館（subject headings）
・国立国会図書館（国立国会図書館件名標目表）

といった各国を代表する図書館がデータセットを公開してきたことが大きな流れをつくっている（括弧内が公開しているデータセット）[12]。また、欧州連合（EU）の国々の図書館・文書館・美術館・博物館の統合サイトであるEuropeanaも実験サイト[13]をつくってLODを指向している。

データの語彙に関する分布は次のようである。Dublin Core（シンプルDC）を使っているデータサイトは全体の約32%、FOAF（Friend-Of-A-Friend）[14]を使っているのは約27%、dcterms[15]が約18%、SKOS（Simple Knowledge Organization System）[16]が約14%である。何らかの独自の語彙も併せて使っているデータセットは全体の約59%、残りは外部で定義された語彙のみで記述している。なお、独自語彙を標準語彙へマッピングする定義が書かれているものは7%程度であった。

先に分野別にデータ数を提示したが、これを各データセットから外へ出ていくリンク（Outlink）数でみると順位は大幅に変わる。生命科学が一番大きくなり50%を超える。以下、出版・論文・図書館関係が約20%、メディア約13%、クロスドメイン約7%である。Outlink数が多いというのはより外のデータとのつながりがあるということであり、生命科学関係はデータセット間でよく参照されているLinked Dataの特徴を生かしたデータであることがわかる。反面、政府関係データセットはデータ数は多いものの、各データセット内で閉じていて、あまりLinked Dataの特徴を生かしていないことを示している。

Outlinkの数でみると、多く（43%）のデータセットは1,000以下である一方、100万を超えるOutlinkを持つデータセットも約11%ある。

一つのデータセットのOutlinkのターゲットのデータセットがいくつあるかをみてみると、ターゲットのデータセットが1つのみが約31%を占める。2つであるのが約19%なのでこれで約半数である。一方、10個以上というデータセットも約14%ある。Linked Dataのサイトといっても多くから参照されているサイトもあれば特定のサイトからしか参照されないサイトもあり、かなり幅があることがわかる。

なお、データセットのうち、データの作成者自身がLODとして公開しているのが約1/3、残りはデータ作成者以外がLOD化している。

4. Linked Dataの役割と期待

データは公開されるだけでも価値があるが、リンクされることによってより価値を高める。これまで各種のデータは紙の文書やPDFで公開されることが多かった。確かに公開はされているが、加工も操作も難しいので、データ提供者の意図どおりに受け入れるしかなかった。データを加工したり他のデータと結びつけたりするという役割はデータの提供者のみに任されていた。

一方、Webページの情報、ことにHTML文書は自由に操作可能である。様々なタギングシステムやリンクシステムでユーザは自分なりの情報のまとめを作ったりすることができる。さらにWeb APIが公開されているサイトではAPIを活用してマッシュアップという形でデータを集約、関連づけることができる。

ユーザがデータを取捨選択できたり他のデータと統合したりできるという点では、Linked Dataの役割はWeb APIと似ている。しかし、Web APIと異なるのはURIとRDFスキーマを用いることで透明性をできる限り確保していることである。透明性があることでデータの統合に関して自由度が増している。このことにより、データの提供者でもデータの利用者でもなく第三者がデータを統合したりすることが可能になった。すなわち、データ提供者以外でも、独自の視点でデータを集約したり加工したりしたデータをまた公開することができる[17]。この点においてはデータ提供者にとってメリットになりうる。すなわち、データ提供者は利用者向けの加工まで用意しなくてもすむようになる。またデータ利用者も好きなデータ加工を選択できるという自由度が得られる。Linked Dataはこのようなデータの利用の役割分担を新たにつくることにより、データ利用をより活性化させることができるのである。

（国立情報学研究所：武田英明（たけだひであき））

（1）Linked Dataについて『情報処理』2011年3月号に特集がある。総説、各分野（メディア、医薬品、政府、地理空間）での状況、日本での課題について個別に言及されているので、こちらも参照されたい。
特集, リンクするデータ（Linked Data）：広がり始めたデータのクラウド. 情報処理. 2011, 52(3), p. 284-333.
（2）Berners-Lee, Tim. “Linked Data”. Design Issues. 2009-06-18. http://www.w3.org/DesignIssues/LinkedData.html, (accessed 2011-05-10).
（3）当初、LODはオープンデータを収集するLinking Open Dataプロジェクトの略称として使われていたが、次第にオープンなLinked Data（Linked Open Data）の略称としても指すようになった。
（4）Cyganiak, Richard et al. “The Linking Open Data cloud diagram”. 2010-09-22. http://lod-cloud.net/, (accessed 2011-05-10).

（5）Cyganiak, Richard et al. “The Linking Open Data cloud diagram”. 2010-09-22.
http://lod-cloud.net/, (accessed 2011-05-10).
（6）Bizer, Christian et al. “State of the Web of Data”, 4th Linked Data on the Web Workshop（LDOW2011）, Hyderabad, India, 2011-03-29.
http://events.linkeddata.org/ldow2011/slides/ldow2011-slides-intro.pdf, (accessed 2011-05-10).
（7）Bizer, Christian et al. “State of the LOD Cloud”. Freie Universität Berlin. 2011-03-28.
http://lod-cloud.net/state/, (accessed 2011-05-10).
（8）Bizer, Christian et al. “State of the LOD Cloud”. Freie Universität Berlin. 2011-03-28. http://lod-cloud.net/state/, (accessed 2011-05-10).
（9）Bizer, Christian et al. “State of the Web of Data”, 4th Linked Data on the Web Workshop（LDOW2011）, Hyderabad, India, 2011-03-29.
http://events.linkeddata.org/ldow2011/slides/ldow2011-slides-intro.pdf, (accessed 2011-05-10).
(10) Bizer, Christian et al. “State of the Web of Data”, 4th Linked Data on the Web Workshop（LDOW2011）, Hyderabad, India, 2011-03-29.
http://events.linkeddata.org/ldow2011/slides/ldow2011-slides-intro.pdf, (accessed 2011-05-10).
(11) Bizer, Christian et al. “State of the LOD Cloud”. Freie Universität Berlin. 2011-03-28.
http://lod-cloud.net/state/, (accessed 2011-05-10).
(12) Bizer, Christian et al. “State of the Web of Data”, 4th Linked Data on the Web Workshop（LDOW2011）, Hyderabad, India, 2011-03-29.
http://events.linkeddata.org/ldow2011/slides/ldow2011-slides-intro.pdf, (accessed 2011-05-10).
(13) Europeana Research Prototype.
http://eculture.cs.vu.nl/europeana/session/search, (accessed 2011-05-10).
(14) FOAF は人と人の関係を書くために定義されたメタデータスキーマであるが、単に人のプロファイルを書くときにもよく用いられる。
FOAF Vocabulary Specification. 2010-08-09.
http://xmlns.com/foaf/spec/, (accessed 2011-05-10).
(15) Dublin Core は 2003 年に ISO 15836 として標準化された（シンプル DC）が、2008 年に提案された豊富で精密な定義をもつ要素に拡張をされた語彙を dcterms と呼んで区別している。
DCMI Metadata Terms. 2010-10-11.
http://dublincore.org/documents/dcmi-terms/, (accessed 2011-05-10).
(16) SKOS はシソーラスや分類表で使われる上位下位関係など概念間の関係を中心とした語彙である。
SKOS Simple Knowledge Organization System Reference. 2009-08-19.
http://www.w3.org/TR/skos-reference/, (accessed 2011-05-10).
(17) 2011 年の東日本大震災における福島第一原子力発電所問題においては、有志が各地の放射線データを集約して公開している元のデータが csv や excel データであるため、工夫して統合しているが、Linked Data であればこういった活動はより楽に行えることが期待できる。
放射線量モニターデータまとめページ. 2011-05-10.
http://sites.google.com/site/radmonitor311/, (参照 2011-05-10).

CA1747

動向レビュー

ONIX：書籍流通における出版社のメタデータ標準化

1. はじめに

書籍をはじめとする図書館資料は、著者による執筆を起点に、利用者がそれを閲覧するまでの一連で流通され、次々に提供される。その中には、出版社、書籍取次、書店、図書館のそれぞれの役割が存在している。各場面において、書籍を流通させ、管理し、探すためには、その書籍を表す何らかのデータ（メタデータ）が必要であることは言うまでもない。

図書館では、書誌データの交換フォーマットである MARC（機械可読目録）が図書館資料の管理および利用者による検索のためのメタデータとして利用されている。MARC は一定の標準規格となっているため、国際的にも多くの図書館で共通で活用することができるようになっている。

一方、出版社、書籍取次、書店の側にもメタデータが必要であることに変わりはない。これまでは各国、各社での独自運用が多かったが、最近では後述する EDItEUR が管理する ONIX というフォーマットの採用が欧米の出版社を中心に進み標準となってきている。

本稿では、書籍流通における商品情報としてのメタデータである ONIX を取り上げ、その概要と共に図書館と関係した動き、および日本での対応状況について解説する。

2. ONIX について

2.1. EDItEUR

ONIX は、出版物の流通における標準化を推進する団体である EDItEUR（European Book Sector Electronic Data Interchange Group）により管理されている。EDItEUR は、1991 年に設立された英国ロンドンに本部を置く国際団体で、19 か国から 80 以上の機関がメンバーとして参加している。日本からは、一般社団法人日本出版インフラセンター（JPO）、株式会社紀伊國屋書店、丸善株式会社の 3 機関がメンバーとなっている(2011 年 4 月 1 日現在)。EDItEUR では、メタデータと各種の識別規格の管理、利用促進を行っており、ONIX 以外にも EDI（電子データ交換による商取引）、RFID（IC タグ）等の標準化、ガイドラインの作成、普及を推進している[(1)]。

2.2. ONIX ファミリー

ONIX は ONline Information eXchange の略称であり、EDItEUR が管理する規格の総称である。それら