Linked Dataを用いた情報統合・活用技術

Information Integration and Utilization Technology Using Linked Data

● 井形伸之　● 西野文人　● 粂　照宣　● 松塚貴英

あらまし

富士通研究所では，Web上でデータを公開する標準的手法であるLinked Dataを用いた情報統合・活用技術の研究開発を行っている。Linked Dataとは，Webの各種標準規格団体であるW3C(World Wide Web Consortium)が推奨するデータ公開のための方法論であり，機械処理を可能とする構造化されたデータ形式が用いられる。近年，学術・政府系を中心に公共性の高いデータがLinked Data形式で公開されており，Linked Open Data(LOD)と呼ばれるWeb上でのグローバルなデータ空間を形成している。

本稿では，著者らがアイルランド国立大学ゴールウェイ校の研究機関 Digital Enterprise Research Instituteと共同で開発した，世界中で公開されているLinked Data形式のデータを収集・格納し，一括検索するLOD活用基盤技術について，活用例を交えながら紹介する。

Abstract

Fujitsu Laboratories Ltd. has been promoting research and development of information integration and utilization technology using “Linked Data,” which is a standard method for publishing data on the Web. Linked Data is recommended by the World Wide Web Consortium (W3C), which is the main international standards organization for the Web. Linked Data uses a machine-readable structured data format that can be processed mechanically. Recently, highly public data such as academic and government related data have been released in Linked Data and created a global data space called “Linked Open Data (LOD)” on the Web. In this paper, we describe our basic LOD utilization technology to collect, store and cross-search worldwide Linked Data with applications. This technology is jointly-developed by the authors and the Digital Enterprise Research Institute of the National University of Ireland, Galway.

ま え が き

近年，著作権などの制限を付けずに，誰でもが利用できるような形でデータを公開する「オープンデータ」活動が活発化している。米国政府は，オープンガバメント政策のもと，政府・公共機関が所有する公共データを一元的に公開するサイト「data.gov」[(1)]を2009年に開設し，2013年5月時点では，米国内の200以上の公共機関から7万を超すデータセットが同サイト上で公開されている。オープンデータ活動の波は世界中に広がっており，現在では40か国以上の政府が専用のデータ公開サイトを開設している。[(2)] 一方，日本では，2012年に内閣府IT総合戦略本部により「電子行政オープンデータ戦略」が策定され，オープンデータの法的な基盤整備や公共データの公開が順次開始された段階にある。

世界各国でオープンデータ活動に取り組む理由として，政府活動の透明性の確保もあるが，公共データの2次利用を促進し，新しい市場を創造する目的もある。EUの研究機関では，公共データを活用したアプリ・サービスの市場規模は280億ユーロ，サービス利用者の効率化や産業競争力強化まで含めた経済波及効果は1400億ユーロと試算しており，これを日本のGDP（国内総生産）比で換算すると，1兆円の市場規模，5.4兆円の経済波及効果と見積もられている。[(3)]

しかし，一口に公共データと言っても，国勢調査のような統計データから航空写真のような画像データまで，様々な種類のデータがある。そのため，上述のデータ公開サイトにおいても，データの種類によって，Excel，JPEG，XMLなど様々なフォーマットが用いられている。また，たとえ同じ種類のデータであっても，データを作成した組織が異なると，異なるフォーマットが用いられる場合もある。

様々なデータフォーマットの混在は，利用者側にフォーマット変換の負担を課し，データの2次利用を妨げる要因にもなる。そのため，データを公開する際には，特定のアプリケーションに依存せず，また機械処理可能なフォーマットが望まれている。

このような中，Webの各種標準規格団体であるW3C（World Wide Web Consortium）では，Web上でデータを公開する手法として「Linked Data」[(4)]を推奨している。

本稿では，初めにLinked Dataの概要や活用事例を概観し，続いて，著者らがアイルランド国立大学ゴールウェイ校の研究機関Digital Enterprise Research Institute と共同で開発した，Linked Data形式で公開されているデータを格納・一括検索する技術について紹介する。

Linked DataとLinked Open Data

● Linked Data

参考文献（5）によると，Linked Dataとは，「Webをグローバルなデータ空間にする仕組み」と紹介されている。現在のWebが，主に「人が読むための文書のWeb」であるのに対し，Linked Dataは「機械処理するためのデータのWeb」と対比される。Linked Dataの基本的な技術要素は，HTTP（Hypertext Transfer Protocol）やURI（Uniform Resource Identifier）といった現在のWebと同様の仕組みを用いているが，RDF（Resource Description Framework）と呼ばれる構造化されたフォーマットでデータを記述するなど，いくつかの違いがある。また，URIは文書単位ではなく，「モノ」単位に振られ，そのデータ構造は，「あるモノ（URI）」に対する「属性名」と「属性値」の組で表現される。例えば，ある人物Aのプロフィール（名前や性別）は，

〈URI〉	〈属性名〉	〈属性値〉
uri：human-A	name	“Alice”
uri：human-A	gender	“female”

のように，「〈URI〉－〈属性名〉－〈属性値〉」の三つの組で記述される。属性値には文字列だけでなく，別のURIを記述することができ，例えば，ある人物Aと人物Bが友達（friend）関係であることは，

〈URI〉	〈属性名〉	〈属性値〉
uri：human-A	friend	uri：human-B
uri：human-B	name	“Bob”

のように記述できる。このようにURI同士の関係を，属性名を使ってひも付けることにより，データはリンク構造を持つことになる。このリンクは，通常のWebと同様に，自分が作成したデータだけ

でなく，ほかの人が作成した異なる種類のデータへも張ることもできる。そのため，Linked Dataで記述されたデータは，データの種別を越えた巨大なネットワークとなる。

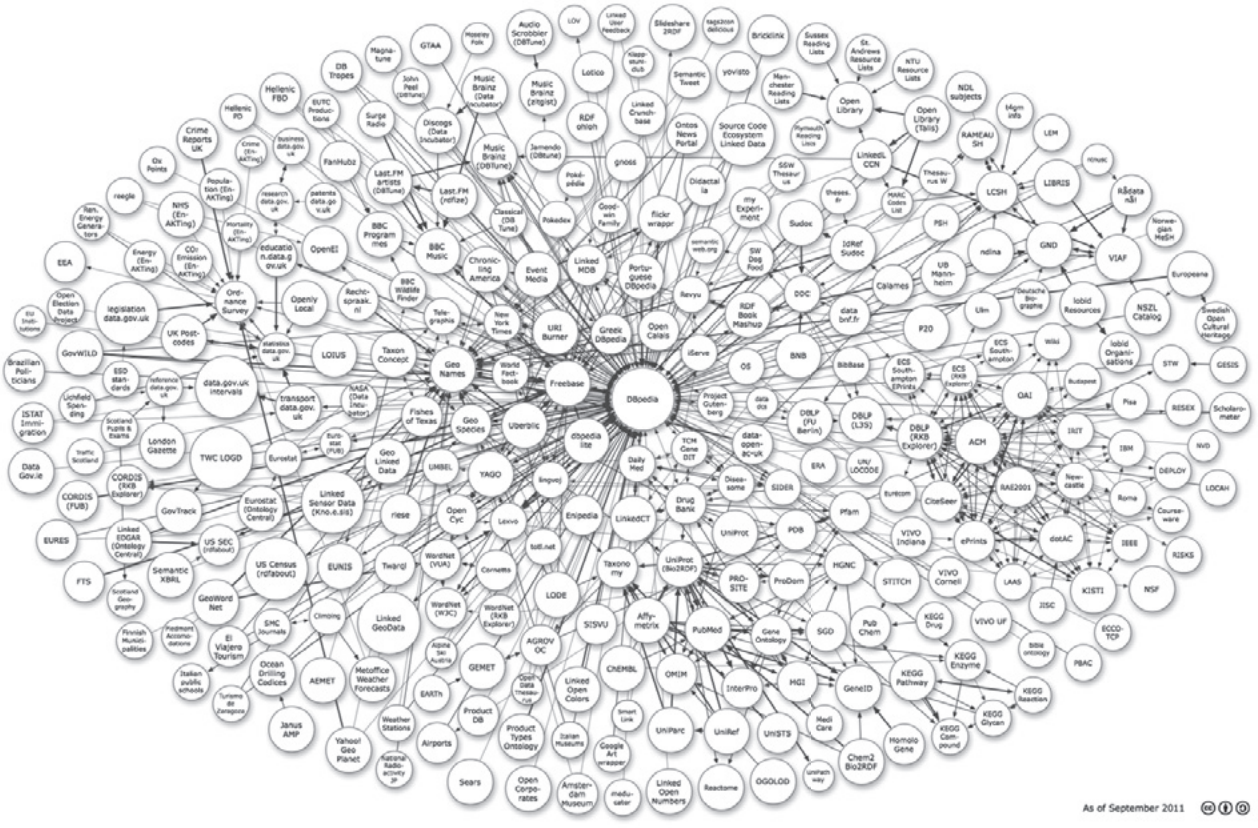
● Linked Open Data

現在，前節で述べたLinked Data形式のデータがWeb上で次々と公開されており，これらのデータ群は，Linked Dataとオープンデータ（Open Data）を合成した「Linked Open Data（以下，LOD）」という言葉で呼ばれ，Web上でグローバルなデータ空間を形成している。

図-1は，現在，Web上で公開されているLODを俯瞰（ふかん）的に図式化したものである。[6] 図中のノードは一つのデータセットを表し，ノード間の矢印がデータセット間のリンク関係を示している。様々な分野のデータセットが公開されており，政府系，地理情報，クロス-ドメイン（共通語彙），ライフサイエンス，学会・図書館系，ソーシャルメディア，マスメディアなどのデータセットがある。代表的なものとして，インターネット百科事典であるWikipediaをLinked Data化した「DBpeida」[7] や英国政府が公開している「Data.gov.uk」[8] などがある。また，日本のデータセットとしては，国立国会図書館が提供する「Web NDL Authorities」[9] がある。2013年3月現在では，LODのデータ量は，全体で400億項目を超える。

Linked Dataの活用事例

Linked Dataは多方面での利用が期待されており，その典型的な活用方法の一つに，コンテンツ・ナレッジマネジメントがある。すなわち，文書，人物，技術概念（キーワード），イベントなど（これらをエンティティと呼ぶ）をデータとして管理し，関連するエンティティ同士をひも付ける（リンクする）ことで，多様な探し方や情報の提供をできるようにするというものである。実際に，富士通内では，技術キーワードなどから関連する技術や関係者，あるいは顧客導入事例などを見つけ



出典：Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/ [6]

図-1　LODの俯瞰図

やすくするために，技術概念，人物，展示会情報，プレスリリース情報，特許・論文，技術キーワード，顧客導入事例などをひも付けたLinked Dataを作成し，ナレッジマネジメントとして活用している。[10]

同様な考えでサービスを提供しているのが，電子情報通信学会（以下，信学会）のI-Scoverシステム[11]である。信学会では，論文誌，研究技報，全国大会論文，国際会議プロシーディングスなどを保有しているが，それぞれのデータに対し別々の検索システムが構築されており，データを横串に検索したり，ある論文から関連する別の論文を見つけたりすることができなかった。そこで，保有する文献，刊行物，人物情報，技術概念情報，イベント情報，機関情報をメタデータとしてLinked Data化し，関連する情報を検索したり見やすく表示したりするシステムを構築した。[12] 信学会では，提供するLinked Dataが情報通信分野のハブとして利用され，今後ページビューが増加することで，学会のプレゼンスが上がり，会員増につながること，あるいは学術・産業界に貢献できることを期待している。

LOD活用基盤

現在，LOD内の各データセットは，データ提供者が個別に立ち上げたWebサイトで公開されている。データを利用する際には，使いたいデータセットを個別に入手して利用することになるが，その際，以下の問題があった。

(1) 欲しいデータがどの公開サイトにあるか分からない。
(2) いくつかの公開サイトではデータファイルが置いてあるのみであり，ダウンロードしてみないとデータの中身を参照できない。

既に使用したいデータセットが特定できている場合は良いが，ある目的に適したデータセットを探したい場合，その手段が存在しなかった。これらの問題に対して，目的に適したデータセットを探すことを可能にし，またデータをダウンロードすることなく，関連データを参照できるようにすることで，LODの利用をより促進することができると考える。

そこで今回，著者らは，アイルランド国立大学ゴールウェイ校の研究機関Digital Enterprise Research Instituteと共同で，世界中で公開されているLODを収集・格納し，複数のデータセットを一括して検索できるLOD活用基盤（**図-2**）を開発した。このLOD活用基盤では，アプリケーション開発者が欲しいデータセットを探すための検索インターフェースを装備し，また，アプリケーションからの利用を想定したデータ問合せ用の標準APIであるSPARQL[13]にも対応している。

このLOD活用基盤により，アプリケーション開発者は，数あるデータ公開サイトを個別に検索したり，データを個別にダウンロードしたりすることなく，必要なデータを一括して入手し利用することが可能となる。また，標準APIを通して，多種

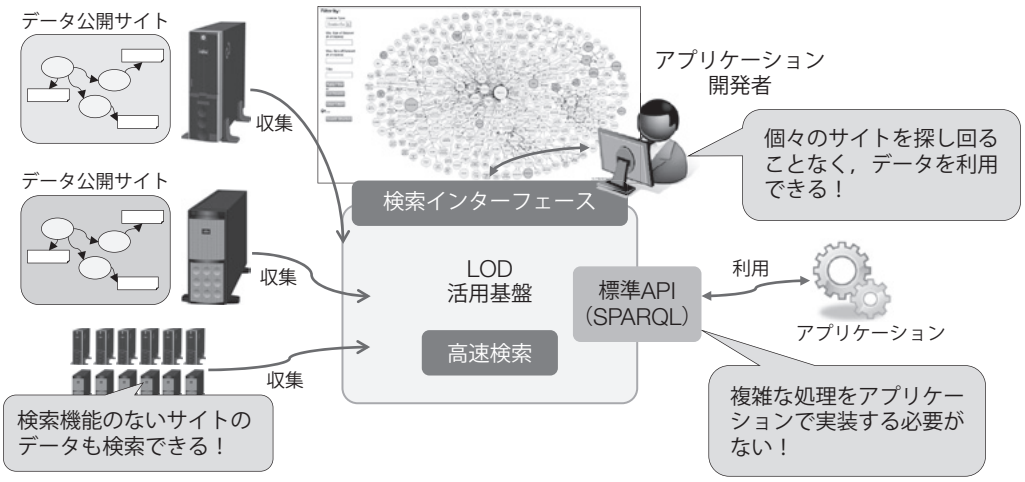


図-2　LOD活用基盤

多様なデータを組み合わせたアプリケーションを簡単に開発することができるようになる。

データを一元的に収集した場合，リンクによって作られる巨大なネットワークを取り扱う必要があり，単にデータ量が増加するだけでなく，複雑化するデータのリンク構造を高速に検索する技術の実現が課題となる。特にデータの中からリンクが張られている共通の項目を検索するような場合，膨大なデータを総当たりで照らし合わせる処理（突合せ処理）が必要となり，これが性能劣化の要因となっていた。

今回，著者らは，このような突合せ処理が必要となる検索処理に対し，Linked Dataに特化した分散処理とキャッシュ機構を組み合わせることにより，従来比5 ～ 10倍の高速化を実現した。

今回開発した検索アルゴリズムの概要を**図-3**に示す。検索条件を調整し，各スレーブサーバで部分的な突合せ処理（1次）を行い，マスタサーバでの突合せ処理（2次）の負荷を軽減できるようにすることで，全体での処理時間を短縮している。また，一部のノードにリンクが集中するといったLODのリンク構造の特徴と過去の利用頻度から，突合せ処理時にアクセスが集中するデータのみを効率的にキャッシュするアルゴリズムにより，ディスクアクセスの回数を抑えることで高速検索を実現した。

LODと公開情報の統合・活用

LOD活用基盤を利用したアプリケーションの具体例として，著者らは，Linked Dataの仕組みを利用して，LODと一般公開情報を組み合わせて，企業を多角的な視点から比較・分析するためのアプリケーションを試作した。具体的には，LOD内のDBpeida[7]やCrunchBase[14]などのデータセットにある企業の基本プロファイル情報と，一般に公開されている各企業の財務報告書や株価情報とを統合・分析している。

通常，これらのデータセットでは異なる企業コードが用いられる。例えば，米国の場合，財務報告書にはCIK（Central Index Key），[15]株価にはTicker Symbol[16]が用いられる。

今回，著者らは，金融分野で標準化が進められているLEI（Legal Entity Identification）[17]をベースとして，一般公開情報をLinked Data化し，LODとの統合（リンク付け）を行った。これにより，企業名や証券コードなどから企業を特定することが可能となるほか，リンクを辿りながら企業に関連する様々な公開情報を芋づる式に参照することができる。

図-4は，試作した企業比較アプリケーションの画面例である。図-4では実在する米国企業に対し，様々なデータセットの情報を軸として比較を行っている。例えば，右下の表では，財務報告書のあ

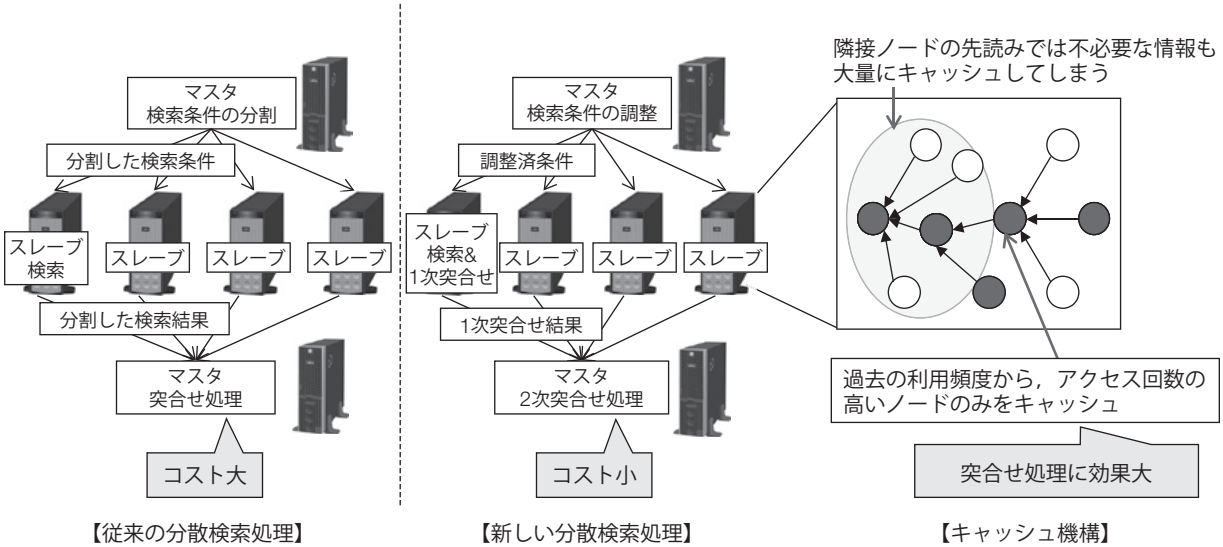


図-3　検索アルゴリズムの概要

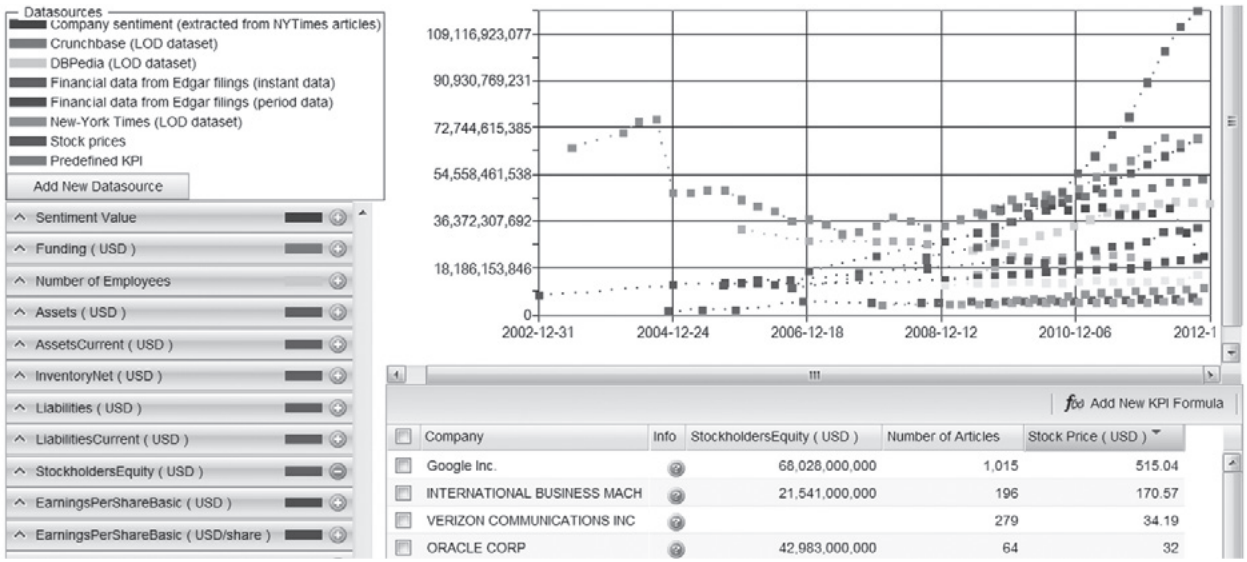


図-4　企業比較アプリケーションの画面例

る財務指標，ある期間中に新聞で取り上げられた記事数，株価情報，といった異なるデータセット中の各数値を横断的に並べて比較している。

このようにLinked Dataを用いることで異なるデータセット間の統合（リンク付け）を簡単に表現でき，また，そのデータを用いて，これまでは難しかった複数のデータセットを横断的に活用した多角的な企業比較・分析が可能となる。

む　す　び

本稿では，世界中で公開されているLODを収集・格納し，一括検索するLOD活用基盤技術について，アプリケーション例を交えて紹介した。

今後，クラウド上に実装したLOD活用基盤を，2013年中に無償公開（限定公開）する予定である。このLOD活用基盤を核として，Linked Dataの普及を加速させるとともに，オープンデータ市場を活性化させたい。

参考文献

（1）米国政府Data.gov.
http://www.data.gov/

（2）Open Data Site.
http://www.data.gov/opendatasites

（3）NTTデータ：オープンデータに関する欧州最新動向．第21回電子行政タスクフォース資料.
http://www.kantei.go.jp/jp/singi/it2/denshigyousei/dai21/siryou1_2.pdf

（4）Linked Data.
http://www.w3.org/DesignIssues/LinkedData.html

（5）トム ヒースほか：Linked Data：Webをグローバルなデータ空間にする仕組み．近代科学社（2013）.

（6）Richard Cyganiak and Anja Jentzsch：Linking Open Data cloud diagram.
http://lod-cloud.net/

（7）DBpedia.
http://dbpedia.org/

（8）英国政府のLOD.
http://data.gov.uk/linked-data

（9）国立国会図書館，Web NDL Authorities.
http://iss.ndl.go.jp/ndla/about/

（10）西野文人ほか：Linked Dataの企業での活用について．情報処理学会研究報告，2011-DD-82（2），p.1-8，2011.

（11）電子情報通信学会：文献検索システム.
http://i-scover.ieice.org

（12）西野文人：I-Scover ～ Linked Dataに基づく電子情報通信学会文献検索システム～．信学通誌，No.25，p.49-53，June 2013.

（13）SPARQL.
http://www.w3.org/TR/sparql11-query/

（14）CrunchBase.
http://www.crunchbase.com/about

（15）CIK：Central Index Key.

http://www.sec.gov/edgar/searchedgar/cik.htm
(16) Ticker Symbol.
http://en.wikipedia.org/wiki/Ticker_symbol
(17) LEI：Legal Entity Identification.
http://en.wikipedia.org/wiki/Legal_Entity_Identification_for_Financial_Contracts

著者紹介

井形伸之（いがた　のぶゆき）
ソフトウェア技術研究所インテリジェントテクノロジ研究部 所属
現在，Linked Dataを用いた情報統合・活用基盤技術の研究に従事。

粂　照宣（くめ　てるのぶ）
ソフトウェア技術研究所インテリジェントテクノロジ研究部 所属
現在，主に知識処理を中心とした研究，およびLinked Dataの活用研究に従事。

西野文人（にしの　ふみひと）
ソフトウェア技術研究所インテリジェントテクノロジ研究部 所属
現在，自然言語処理・知識処理，およびLinked Dataを活用したナレッジマネジメント，コンテンツマネジメントシステムの研究に従事。

松塚貴英（まつつか　たかひで）
欧州富士通研究所 所属
現在，Linked Dataの大規模分散データベースおよび処理基盤の研究に従事。