Linked Open Data 調査報告

セマンティックWeb委員会
○長野 伸一　株式会社 東芝
清水 昇　慶応大SFC研究所（(株)サイバーエッヂ）
高島 周二　株式会社 サイバーエッヂ
細見 格　日本電気株式会社
佐藤 宏之　日本電信電話株式会社
飯塚 京士　日本電信電話株式会社
津田 宏　株式会社 富士通研究所
乙守 信行　株式会社 MetaMoJi

2010年03月05日

Linked Open Data（LOD）

- セマンティックWebにおける新しい潮流
  - 欧米を中心に，W3Cや分野のコミュニティにより，データの公開と利用が進められている
- 事実（fact）主体のデータセット
  - データ同士がリンクされ，大きなネットワークを形成．データのクラウドとも呼ばれている
- LODを利用した事例が増えつつある
  - 特定の分野内だけでなく，分野を横断した利用も
  - 企業によるLODの利用の試みも始まっている

調査方針

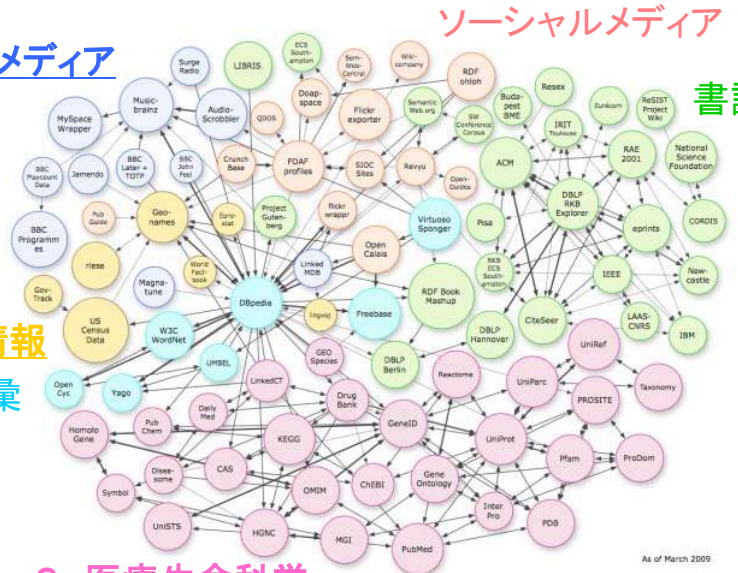
- 目的指向で取り組んでいる事例を対象として，以下の観点で調査を実施
  - 解決したい課題は何か（ビジネス／利用者の視点で）
  - LODを利用して，どのように解決しようとしているのか
  - 狙っている効果は何か
- 調査の対象分野
  1. メディア
  2. 医療生命科学
  3. 政府
  4. 地理情報

調査対象分野

ソーシャルメディア

1．メディア

書誌



3．政府

4．地理情報

共通語彙

2．医療生命科学

http://linkeddata.org/

1．メディア：BBC放送

BBC放送が抱えていた課題

1.8つのTVチャンネルと地方局、10個のラジオ放送局、また40の地方ラジオが各サイトが個別に管理されてきた

2.ある番組のプログラムで紹介されているアーティストを辿ることさえできない状態

3.毎日1000から1500の番組を人手でカバーすることは無理な状況

4.BBCはLinked Open Dataによるコンテツ連携を開始した

BBC放送が目指すサイト

1.ユーザビリティの向上とユーザ体験の向上
利用者の興味や関心に対応サイトの構築

2.ユーザーがBBCコンテンツで旅をする
BBC/natureでは生息地などの関連づける

3.Webサイト自体がAPIとなる
ひとつのURIで管理。Webサイトはサードパーティも利用可能

4..緩やかな関係に基づく開発を可能とする

導入技術ポイント1

BBCの各サイトの情報をDBpediaを統制語彙とし、またBBC programmes オントロジーで関係づける

導入技術ポイント2

BBCで保持していないコンテンツを Linked Open Dataで各番組に関係づける

5

1．メディア：BBC放送

BBCのサービスと背景の技術：　BBC Programmes

サービス

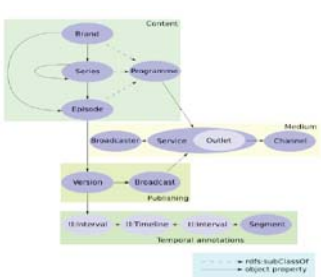
BBCのTV,ラジオのプログラムのポータルサイトであり、2007年夏より開始される。

1ページが1番組のパーマネントURIで管理され、また各番組のページはBBC Programmes オントロジによって関係付けられる。

BBC Programmesオントロジ

BBCが元々所持していたDBをベースに作成したオントロジー

Brand(番組分類)、Series(番組のシリーズ),Episode(番組放映分)とスケジュールなどをオントロジーで管理することで、カスタマイズ性が高い番組ページの開発が可能となった。



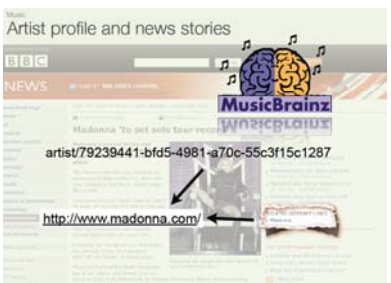
6

1．メディア：BBC放送

BBCのサービスと背景の技術：　BBC Music

サービス
アーテイスト紹介ページ(Music)から、アーティストの履歴や関連サイト紹介など利用者の興味に合わせた多種な情報で「旅」を出来るサービスを実現する。

CMS（Content Management System）としてのWebページへ
BBCがコンテンツとして保有していない情報についてLinked Open Dataを利用してサイトに関係付ける。アーティスト紹介のページでは、MusicBrainz/Wikipediaなど
このときDBpediaを統制語彙として利用



7

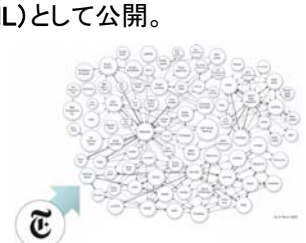
1．メディアのLODに関連する取り組み

New York Times（NYT）

2009年6月26日、LODクラウドへの参加を発表

- 過去150年間の新聞記事用語 100万語以上を5つのカテゴリ（記事の題名、個人名、組織名、地理名、作品名）に分類して公開していく。
- 新聞業界の中でいち早くオンライン事業を展開。紙媒体に代わり、電子媒体を新たな収益（オンライン広告など）基盤としたいと考えている。
- Times Open戦略：新聞の使命に基づき、オープンなインターネットの基盤の上で、ソーシャルネットワーキングと集合知を活用し、ジャーナリズムとして行動する。
- Times Tags API で既に提供中の**27,000語**に加え，用語を順次公開していく。
- 2009年10月29日、2010年1月13日の2回に分け，クリエイティブ・コモンズのライセンス条件のもとに、**10,000語**を**LOD（RDF/XML）**として公開。

LOD化作業スケジュール
第1段階　1980年から現在までの記事から数十万語
第2段階　1851年から1980年までの記事から数十万語



1. メディアのLODに関連する取り組み

Thomson Reuters

2009年1月、Calais4.0ウェブサービス及びそのOpen APIを出版社向けに公開する

- Calais：人や場所、会社、事実やイベントなどにメタタグを付与し、様々なコンテンツをリンクさせるウェブサービス。出版社のもつコンテンツからWikipedia, DBpedia, GeoNames, the Internet Movie DataBase(IMDB), Shopping.comにリンクすることが出来る。

CNET, Huffington, DailyMe

2009年6月Calaisを採用

- Huffington Post ：著名ジャーナリストの政治ブログから始まって拡大し、今では様々なジャンルにわたってニュースを提供している巨大サイト
- DailyMe :MITメディアラボによる、好きなカテゴリなどを選べば自動でニュースが配信される、パーソナライズされた新聞

Harpers Magazine

セマンティックWeb技術の導入を進めている（LODの公開は行っていない）。

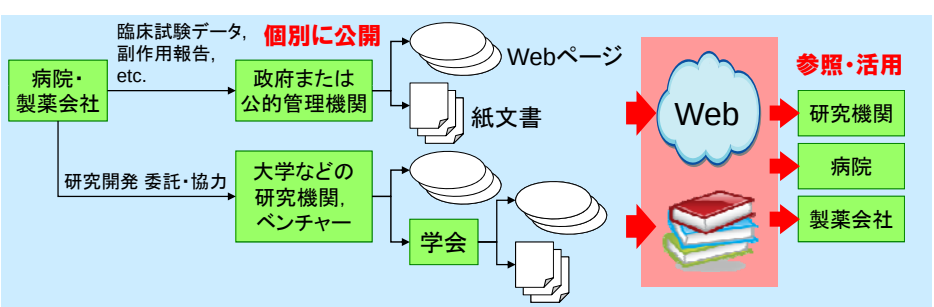
- USで最も古い雑誌出版社のひとつ。と同時に現在は、技術的に最も進んでいる会社となっている。200,000の個人法人を有している。
- Weekly Reviewの2000年からのデータコンテンツがあり1850年に遡った（創業時）フルコンテキストのスキャンを雑誌社が始まった時から行っている
- 2003年からセマンティックWeb技術の導入を開始。コンテンツをカテゴリにセグメントすることからタクソノミー化し、オントロジーにより小さいセクションにわけたコンテンツを関係付ける。
- 300ページの静的なページをつくるために1100ページの再合成したページを利用することにより有効性が劇的に向上、Webサイトのトラフィックが増える、サブスクリプションの収入が増える、Webサイトのメンテナンスコストが下がるなどの成果を得ている。

9

2. 医療生命科学

治療・健康維持・創薬などを目的とした学問領域

- 病院・製薬会社・研究機関の他、伝染病や高齢化社会などへの対策で政府の関与も大きい
- 大学や学会から調査・研究成果が、（欧米）政府からは薬品等の臨床試験データが公開



2. 医療生命科学における課題

テーラーメイド医療による副作用/医療費削減

- 医薬品副作用による医療費増大（日本で1.7兆円@2008年）
- 遺伝子解析の進展等から、患者の体質に合った処方に期待

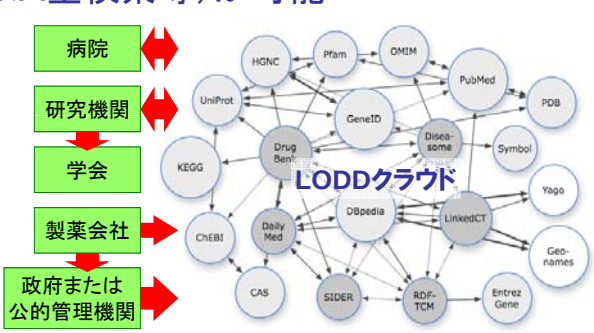
- 遺伝子、たんぱく質、代謝、病気、薬剤、患者等の多岐に渡る情報を統合した分析が必要
- 情報は豊富だが、収集・活用は容易でない
  - 異なる情報源：互いに独立し、関連情報の発見困難
  - 異なるアクセス手段：個別に習得や実装が必要
  - 文書やページ単位の検索：欲しい情報を見つけ難い

2. Linking Open Drug Data（LODD）

W3Cの医療生命科学分野に関するプロジェクト

- W3C Semantic Web Health Care and Life Sciences Interest Groupの活動，製薬会社（Johnson & Johnson, etc.）など約30団体が参加
- 薬剤情報を中心としたLODDクラウドを形成
- http://esw.w3.org/topic/HCLSIG/LODD

- 20以上のLODが相互にリンク，関連情報の素早い参照や計算機処理（Q&A型検索等）が可能
  - 薬剤，副作用事例，
  - 病気，臨床試験，
  - たんぱく質，遺伝子，
  - 医療保険制度，
  - etc.
- 用語の共通化・対応付け
  - LODは事実（fact）の集合
  - 概念や用語の対応付けはSNOMED-CTなどの医療オントロジを活用



3. 政府データのLODに関連する取り組み

政府のデータ公開やそれに呼応したもの

- data.gov
  - 米国連邦政府のさまざまな機関が扱う情報を入手できるサイト（2009年5月開設）
  - オバマ政権の開かれた政府（Open Government実現手段の一つとして注目を集める
- The Data-gov Wiki
  - 上記data.govのデータセットをRDF化して公開
- data.gov.uk
  - 英国政府の取り組み（2010年1月開設）
  - データ公開・再利用に向けて、セマンティックWebを意識したサイトになっている
    - Tim Berners-Lee卿、W3CのeGovernmentの取り組みを進めるJohn Sheridan氏のアドバイスを取り入れている

市民活動によるもの

- GovTrack.us
  - 米国議会に関するデータを公開
- The 2000 U.S. Census: 1 Billion RDF Triples
  - Jashua Tauberer氏による米国国勢調査データセットのRDF化

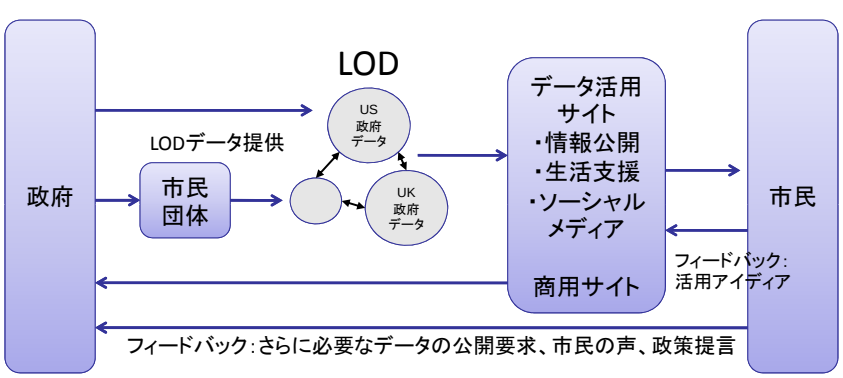
※政府の透明性の向上と市民の啓発を目的とした取り組み

いずれも、広く参照してもらったり、再利用されたり、活用されたりするためのツールとして，データのLOD化が期待されている

13

3. 政府データの活用における課題

単なるデータ公開だけでなく、Linked Open Dataの仕組みを用いて、Open Governmentの「透明性」「国民参加型」「協業的」といった理念を実現できるか？



14

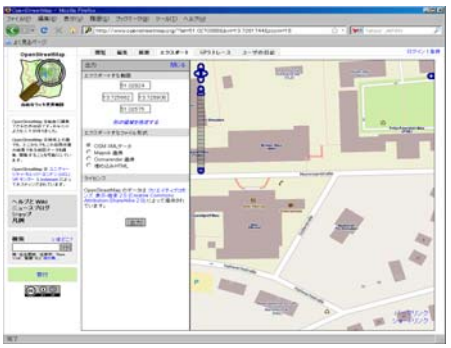
3. 政府データの活用に向けた現状

- データを活用したアプリケーションを探る試み
  - Data.govを活用したアプリを生成するイベント（Semantic Hackathon）などが開催されている
    - Web2.0的なデータ加工を行い活用方法を模索している段階
      - 統計情報を分かりやすく可視化
      - 地図上での各種データのマッシュアップなど
  - data.gov.ukでは賞金を出して活用アイディアを募集
- 商用利用も奨励
  - data.gov.ukの公開情報に基づいて、ケアハウスを比較して選べるサイト（Best Care Home）が立ち上がり、ケアハウス側からも有料で掲載して欲しい情報を提供できるようにすることも検討されている
    - 今後、こうした試みがLODベースで進む可能性がある

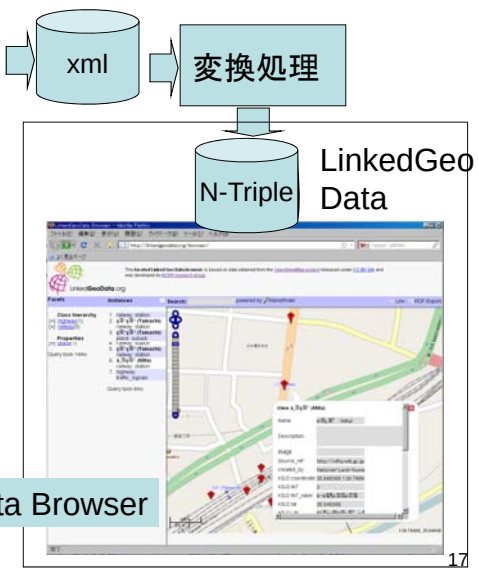
4. 地理情報：Linked Geo Data（LGD）

- http://linkedgeodata.org/ （独ライプニッツ大)
- 概要
  - Webに空間的な次元のオープンなセマンティック情報を付与する為の活動。
  - OpenStreetMapプロジェクトにより収集された情報を活用し、Linked Data原理に則ったりRDF知識ベースとして利用可能にする。これにより、Linked Open Dataとして公開されている他の知識ベースとの相互連携を実現することを狙いとする。
  - OpenStreetMap (http://www.openstreetmap.org/)：誰もが利用できるフリーの地理情報データの開発プロジェクト
- サイズ
  - 形態：N-triples形式のものをダウンロード可能(RDF/XML形式では公開なし)
  - 内容：欧米の情報が主体、日本に関する情報は少ない。
  - 量 ：大量 LOD2クラウドで最大のデータ(20億トリプル)
    - ①3億5,000万箇所(nodes)の情報
    - ②3,000万の道路(ways)の情報
  - 公開方法
    - ①ダウンロード
    - ②専用ブラウズィングページの公開
    - LGDブラウザ http://linkedgeodata.org/browser/ （日本語には未対応)

4. OpenStreetMapとLGDとの関係



OpenStreetMapのデータのexport画面



Linked Geo Data Browser

4. 地理情報データ:他の事例

GeoNames

- http://www.geonames.org/
- 国、大都市、首都、山、郵便番号などの地名(800万以上)と、緯度経度高度、人口などの関連をRDF化。
- スイスのソフトウェアエンジニアである、Marc Wickが立ち上げ。Creative Commonsライセンスで公開。
- National Geospatial-Intelligence Agency‘s (NGA)、U.S. Geological Survey Geographic Names Information Systemなど多くの公開データから作成

DBpedia Mobile

- http://wiki.dbpedia.org/DBpediaMobile
- 地理関係のLODを利用したアプリの一つ。現在地(GPS)から、関連するDBpediaなどの情報を地図上にリンク
  - 218万のthingsのうち、場所として約30万が登録
  - GeoNames, Revyu, EuroStat, Flickrなどの情報の組合わせ
  - Firefox 2, Internet Explorer 7, Safari 3 and Opera 9、モバイルではOpera Mobile8に対応

4. 地理情報データの課題

- 地理情報については、巷に立ち上がっているサービスや公的機関(省庁など)の公開情報と比較して、ビジネスモデルがはっきりしない
- 日本(語)の対応遅れ
  - GeoNamesでは日本は27,220箇所登録されており、51位。
    - 最も高い山が富士山でなかったり、埼玉の県庁所在地が浦和だったり、東京都庁はあっても滋賀県庁はない、とか多少データは怪しい。
  - LGDブラウザでは日本語が正しく表示されない
- LGDの公開データ構造として何が望ましい?
  - RDF/XML, OWL, N-triples, N3
- 情報の信頼度・鮮度・カバレージがまちまち
- どのように情報を更新するか:変化の多い地理情報に対して、GoogleMapsやセカイカメラなどコンシューマを巻き込んだようなサービスにするか、あるいはYahooディレクトリ, GeoNamesのように熱意のある少数編集者による更新モデルか。

まとめ

今回調査した範囲で分かったこと

- 相互にデータを公開・利用することで，互いの問題を解決できる効果が期待できる
  - 現実の課題を共有できれば，LODの公開と利用の目的が明確になる（メディア，医療生命科学）
  - インフラとなるデータは公開の一方向となる．更新の動機付けが必要（政府，地理情報）
- ベストプラクティスと呼べる事例はこれから
  - 効果，有用性の検証を進める
  - ビジネスモデルの確立に向けてトライアルを
  - 情報の信頼度・鮮度・カバレージを定量化し，用途・目的に応じて利用者が選べるように
- 日本語データの整備を進め，世界に向けて発信を

参考資料

- W3C SWEO Linking Open Data Project
  http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData/
- 上記Webサイトの日本語訳
  http://www.semanticweb.jp/lod/SWEO.html