3E3-OS20-9

日本における Linked Data の普及にむけて

Challenges for Linked Data in Japan

武田 英明[*1*2] Hideaki Takeda
嘉村 哲郎[*2] Tetsuro Kamura
加藤 文彦[*1] Fumihiro Kato
大向 一輝[*1*2] Ikki Ohmukai
高橋 徹[*3] Toru Takahashi
上田 洋[*3] Hiroshi Ueda

[*1] 国立情報学研究所 National Institute of Informatics
[*2] 総合研究大学院大学 Graduate University for Advanced Studies
[*3] ATR プロモーションズ ATR Promotions

In this paper, we introduce our project called LODAC which aims to build Linked Data infrastructure in Japan. Linked Data approach is beneficial for data publishers in three folds; providing standardized publishing method, providing integration method for distributed data, and providing the method for inter-disciplinary use of data. We are currently working with data in Museum area and Geo area. In Museum area, we build a service called LODAC Museum where data of museum collections and creators are accumulated, integrated, and published as Linked Data. Creators of the collections are identified by referring the art thesaurus and associated to individual works. In Geo area, address data and public facility data are published with links to each together as Linked Data.

1. はじめに

Linked Data は新しい情報公開・共有の仕組みとしてヨーロッパおよび米国で認知されつつある。本稿では日本における Linked Data の普及に向けた我々のグループの活動の概要を説明する。なお、Linked Data の社会的価値や普及に向けての問題点などは[武田 2011]を参照されたい。

2. LODAC プロジェクト

我々は 2010 年 4 月より「学術リソースのためのオープン・ソーシャル・セマンティック Web 基盤」（通称 LODAC）プロジェクトを開始した。国立情報学研究所が所属する情報・システム研究機構は4つの研究所からなるが、研究所分野を超えた研究を活発化させるために新領域融合研究センターというものを設け、学際的な研究を推進している。LODAC プロジェクトは、そこで行われているプロジェクト（異分野研究資源共有・協働基盤の構築）の中の1つのサブプロジェクトである。

LODAC プロジェクトでは、広く学術に関する情報・データを共有する仕組みを Linked Data の方法で構築するということを目標としている。このプロジェクトでは単に情報基盤を構築して分野の研究者やコミュニティに供給するというのではなくて、当該分野の研究者と一緒となって実際に使える形のサービスを作るところまでを実験的に行うことを狙っている。

3. Linked Data のメリット

Linked Data を用いることのメリットは 3 段階に分けることができる。

(1) オープンデータ化の手段

Linked Data はデータを公開して共有する手段として有効である。URI を識別子として使うことで、データ内の個別の事項をグローバルに一意的に指すことができる。また標準的な方法でアクセスすることができる。またデータの表現が RDF という一つの方法であるので、データ処理において単純化される。

Web でデータ公開をしていると言っても実際には HTML 文書、PDF 文書、csv など様々な方法がとられている。これらに比べ、Linked Data によるデータ公開はアクセスの容易性やデータの再利用性で優れている。

(2) 分野内データ共有の手段

一つの分野内でもデータを公開しているサイトは一つであるとうことは稀であろう。むしろ、複数のサイトが各々のもつデータを公開するというのが普通である。このような分散的な情報を統合的に扱うときにも Linked Data は効果がある。

(1)で述べたようにデータのアクセス手段・表現手段が統一されていることに加え、データサイトが相互にリンクを張ることにより関係性も表現できる。またデータ構造を規定するスキーマの対応も表現できるので、類似のデータ構造であれば、その違いを吸収してアクセスすることも可能である。

データをどこかに集約するのではなく、分散されたデータを分散しつつ、統合的なアクセス・利用を可能にするという点で Linked Data は有効な方法である。

(3) 分野を超えたデータ共有

さらに分野を超えたデータ利用にも Linked Data は効果がある。一般に他分野のデータはアクセス手段やデータ構造がわからず利用が難しい。Linked Data であればアクセス手段は一定のものが保証される上、公開されたスキーマはオントロジーとして表現することができるので、他分野のスキーマを理解して利用することが容易になることが期待できる。
分野を超えたデータ利用はデータに新しい価値を付加する。これがまたデータを公開することのメリットとなる。

我々はこのような点を念頭に置いて、現在いくつかの分野でデータの Linked Data 化を行っている。

連絡先：武田英明，国立情報学研究所，千代田区一ツ橋 2-1-2，takeda@nii.ac.jp

4. LODAC Museum

最初に注目したのは美術館・博物館情報である。美術館・博物館の情報は一部の集約サイト（文化遺産オンライン http://bunka.nii.ac.jp/、国立美術館所蔵作品総合目録検索システム http://search.artmuseums.go.jp/）をのぞけば、館ごとに情報公開がなされており、きわめて分散的である。従って(2)のメリットが期待できる。LODAC Museum ではこの分散された情報を統合的に利用できるサービスを提供することを目的として構築を始めた[嘉村 2010]。

4.1 LODAC Museum の情報源

日本には 6000 館に近い美術館・博物館が存在するが、Web により収蔵品や展示品の情報を公開している館はそれほど多くない。また公開している場合でも HTML ページとして公開しており、データの共有や利用においては問題がある。

本来は情報を公開する館が Linked Data として公開するのが望ましいが、今回は我々がデータを scraping し、Linked Data に変換している。現在、14 館の情報を収集している。他にも国指定文化財データベースなども用いている。また Wikipedia を簡易的に Linked Data 化する dbpedialite[1]を通じて日本語 Wikipedia の情報も利用している。

4.2 データの表現と統合の方法

このような複数の情報源からの情報を扱うと、どう統合するかが問題になる。ここでは次のような方針を採った。

(1) 標準的なスキーマによるデータ表現

美術品のメタデータは詳細度の程度の差があるものの、表現内容はそう大きく変わるものではない。そこで、今回対象にした館での美術品のメタデータをおおよそカバーする程度のメタデータを定義し、個々の館からのデータをこれに当てはめた。

(2) 情報源ごとの ID と統合 ID

個別の館からのデータ（作品、作者）にはそれぞれに ID をつける。一方、それとは独立に統合 ID を個々のデータに振る。もし、一つの作品データや作者データが二つ以上の情報源からくる場合は、一つの統合 ID に複数の情報源ごとの ID が関係づけられる。

(3) 統合のためのシソーラス

上記のような統合をするときには信頼できるシソーラスが重要になる。今回は日本美術シソーラス[福田 1997]を利用して作者情報の統合を行った。

4.3 実装状況

RDF データベースとしては BigOWLIM[2]を用い、SPARQL Endpoint (http://lod.ac/sparql)と簡易ブラウジングインタフェース(http://lod.ac/)を用意している。現在、約 10 万のデータ（作品、作者等）が格納されている。統合結果としては、日本美術シソーラスにある 1332 人のうち 615 人が作品データに使われ、作品数の約 1/4(15020/61861)をカバーしている。

今後はさらに収集館を広げると共に統合作業を進める予定である。

5. その他の分野

5.1 地理・地図情報

地理・地図情報はデータを共有する軸として良く使われる情報の一つである。LODAC Museum には美術館・博物館施設の情報が含まれている。それを他分野のデータと容易に結び付けられるようにするために、より汎用的な地理・地図情報の Linked Data 化を試みている。

最初に注目したのは住所情報である。住所情報として、大字・町丁目レベル位置参照情報 (http://nlftp.mlit.go.jp/isj/) と、郵便番号データ (http://www.post.japanpost.jp/zipcode/) の Linked Data を試作した。また、汎用の施設情報として、国土数値情報施設情報 (http://nlftp.mlit.go.jp/ksj/) と駅データ (http://www.ekidata.jp/) を Linked Data 化し、それを住所情報に結びつけることも試行している。現在 http://location.lod.ac/ にて試験公開をしているが、今後は LODAC Museum と統合する予定である。これにより、地理・地図情報を通して美術館・博物館とは直接関係のないデータが繋がっていくことを期待している。

5.2 イベント情報

横浜 LOD プロジェクト[3]に協力して、地域イベントの Linked Data 化も試行している。イベント情報には地理情報、出演者・出展者といった作者情報が含まれる。これらを先の美術館・博物館情報や地理情報と結びつけることで、分野を超えたデータ利用のユースケースを実現することを狙っている。

6. まとめ

本稿では現在我々のプロジェクトで行っている活動を紹介した。日本においてセマンティック Web の知名度の低さやそもそもデータのオープン化に対する意識の低さからなかなか情報提供者自身がデータを Linked Data 化するというところまでいたっていない。このため、本プロジェクトではそこからはじめ、Linked Data のメリットを実際に見える形で提示すべく活動を行っている。

一方、日本でも徐々にオープンガバメントに関する関心が高まっている。オープンガバメントは米英では Linked Data と親和性が高いと認識されている。このような動きにも注目しつつ、今後も Linked Data の日本での普及に貢献できるよう活動するつもりである。

参考文献

[武田 2011] 武田英明, 日本における Linked Data の現状と普及に向けた課題, 情報処理, Vol. 52, No. 3, pp. 326-333 (2011)

[嘉村 2010]嘉村哲郎, 加藤文彦, 大向一輝, 武田英明, 高橋徹, 上田洋:Linked Open Data による多様なミュージアム情報の統合, 人文科学とコンピュータシンポジウム論文集, 情報処理学会シンポジウムシリーズ, 第 2010 巻, pp. 77-84 (2010).

[福田 1997] 福田 博同, 五十殿 利治:美術シソーラスデータベース形成の諸問題,情報管理, Vol. 40, No. 9, pp. 790-809 (1997).

[1] http://dbpedialite.org/

[2] http://www.ontotext.com/owlim

[3] http://scholex.com/ocdi/