

日本語 Linked Data に基づくモビリティサービスの実現

A Mobility Service based on Japanese Linked Data

飯島千絵
Chie IJIMA

森田武史
Takeshi MORITA

榎本善太郎
Yoshitaro ENOMOTO

山口高平
Takahira YAMAGUCHI

慶應義塾大学
Keio University

The objective of the present study is to implement a mobility service based on Japanese Linked Data. We focus on the modifiability of data, the frequency of update for the data on the websites, and separate the treatments for the datasets depend on the degree of frequency of update to show the possibilities for management of data as Linked Data that consists of various kinds of data. As a practical matter, we created Linked Data in Japanese focused on geographical or positional data and developed a mobility service application to support driving and planning to go for a drive using that based on it.

1. はじめに

近年、米国・英国において、政府所有データが RDF 形式で Web 上への公開が始められている。一方、このような動きが後押しする形で、複数のオープンデータを RDF で表現し、相互にリンクづけられたデータである「Linked Open Data(LOD)」が注目されており、情報流通基盤として期待されている。現状では、英語のデータが中心であり、日本をはじめ、アジア圏では、まだ Linked Data として公開されているデータは少ない。また、情報サービスへの適用方法はまだ十分に研究されていない。

本研究では、地理・位置データを中心とした日本語の Linked Data を構築し、それを利用した運転走行支援およびドライブプラン作成支援のためのモビリティサービスを実現することを目的とする。

データの更新性に着目し、更新性の高いデータと低いデータで処理を分けることにより、様々なデータを Linked Data として扱うことができる可能性および Linked Data の情報サービスへの適用可能性を示す。

2. 関連研究

DBpedia Mobile [Becker 08]は、前述の DBpedia の持つ位置データを中心としたデータを地図上にマッピングするアプリケーションである。モバイルデバイスの GPS が取得した現在地を基に、その周辺の位置情報を持つ、データを地図上にマップして表示する。それぞれのデータの詳細情報を持つ、RDF データは、Marbles Linked Data Browser で閲覧することが可能である。

ユーザは、その位置の背景情報を知ることができ、当該データとリンクしているデータを、データ間のリンクを通じて得ることができる。DBpedia 以外のデータセットとしては、GeoNames, Revyu, EuroStat, Flickr が利用可能である。これらのデータセットは、比較的更新性の低いデータである。本研究では、更新性の高いデータも扱い、Linked Data として情報サービスに利用していく可能性を検討する。

3. 提案システム

3.1 システム概要

Web 上には、様々な種類のデータが混在している。それらを分類するには、例えば、構造度・信頼性・更新性の 3 つの軸が考えられる。本研究では、データの更新性に着目する。Linked Open Data には、更新性の低いデータが多いが、情報サービスを開発するにあたっては、更新性の高いデータを利用することも多い。本研究では、それぞれのデータの更新性の度合いに応じて、最終的にアプリケーションから Linked Data を利用する方法を分離する。更新性の低いデータは、データベースに格納して利用し、更新性の高いデータは、リンクを辿り、API 形式でデータを取得して利用する。図 1 にシステム構成図を示す。

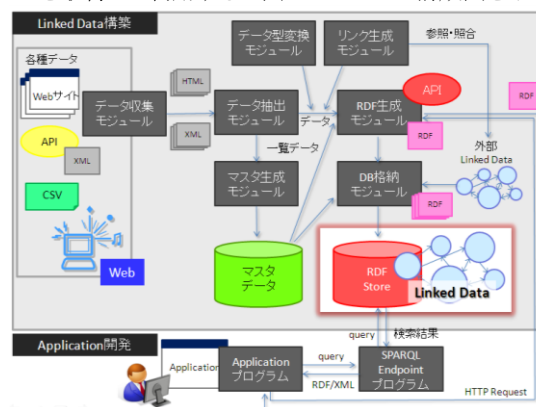


図 1 システム構成図

3.2 Linked Data 構築およびアプリケーション開発方法

(1) データ収集・マスタ生成

アプリケーションのドメインに必要なデータを持つ、HTML・XML・CSV 等のデータを収集する。

データ抽出にあたっては、それぞれの Web サイトの HTML の構造は大きく異なっているため、それぞれに対応するスクレイピングプログラムを開発し、データ抽出を行う。必要に応じてマスタデータを生成しておき、RDF 変換や、データベースを格納

する際に使用する。

(2) RDF 変換・データ型変換

抽出したデータに対してモデリングを行い、RDF 形式に変換する。本研究では、RDF 形式に変換するにあたっては、PHP 言語で RDF データを扱うためのライブラリである、RAP (RDF API for PHP V0.9.6) を用いてそれぞれのデータをモデリングに合う形になるように RDF のステートメントを記述していくことにする。RDF 変換モジュールは、リソースの URI に HTTP Request を送ると、そのリソースの情報を表す、RDF データを返すプログラムとし、最終的にそれを API として利用し、データベースに格納する際もその API 経由で格納できるようにする。

RDF の記述の際、必要に応じて時間や日にちの表記は、値のデータ型を変換し、全て統一してから記述する。本研究では、SPARQL に対応している、XMLSchema の標準的なデータ型に統一して記述する。

(3) リンク生成

RDF に変換したデータセットのリソース間や、外部の Linked Data ヘリンクを生成する。

本研究においては、正式名称があり、それにより記述されているリソースの場合は、リソースの名前の文字列の完全照合により 2 つのリソースが同義であるかを判断する。しかし、店舗等のデータの場合、名称に表記ゆれ等を多く含んでいて、かつ位置データの場合は、以下の 2 つ基準を満たすかどうかにより同義であるかを判断することにする。前提として、リソースは、緯度・経度、または住所の情報を保持しているものとする。

(1)住所が一致する/それぞれの地点間の距離が一定基準内である

住所のデータをそれぞれが持っている場合は、住所が一致することが条件となる。住所のデータがなく、緯度・経度の値のみを持っている場合、同じ店舗のことを表現している、緯度・経度の値が全く一緒である場合は少ない。そのため、各リソースの緯度・経度から求めた距離が、一定基準内(300メートル以内等)であるかを判断する。

(2)地点の名称の文字列に部分照合する

表記ゆれに対応するため、それぞれの名称の、空白や「・」を取り除いた上で、文字列を比較し、照合するかどうかを判断する。

また、本研究では、位置データを中心とした Linked Data を構築するため、地点を表すデータには、行政区画 RDF データを同様の手法で用意し、そのデータへのリンクを付与する。これは、地域情報へのリンクのハブになることを目的としている。行政区画 RDF データは、日本郵便の提供している郵便番号データを利用して、都道府県→市町村→大字といった階層構造を表現する。

(4) DB 格納および SPARQLEndpoint の作成

更新性の低いデータは、RDF データをデータベースに格納する。本研究では、MySQL のデータベースに、PHP のライブラリである ARC2 を用いて、格納することにする。データ抽出の際に用意した、マスタデータを利用し、インスタンスの一覧を取得し、すべてのインスタンスの URI に対して HTTP リクエストを送ることで、API 形式になっている RDF 変換モジュールより、すべてのインスタンスの RDF データを取得し、順に格納していく。外部の Linked Data も合わせて使う場合には、そのダンプデータや、SPARQL の Endpoint・公開されている API 等を利用し、合わせてデータベースに格納する。

さらに、そのデータベースに対して、検索をかけることができるようにするために、SPARQL Endpoint を作成する。今回は、SPARQL の SELECT および CONSTRUCT クエリに対する結果を返すことができるプログラムを Endpoint とする。

(5) アプリケーション開発

以上のプロセスから Linked Data を構築し、それを利用してアプリケーションを開発する。更新性の低いデータは、Endpoint を利用し、SPARQL クエリを投げ、更新性の高いデータは、API を利用し、RDF データを取得する。

3.3 構築した Linked Data

今回データとして、高速道路の関連するデータとして、高速道路・サービスエリア(SA)・インターチェンジ(IC)、さらに、レストランレビューデータ、地域情報として、イベント情報・天気予報情報、交通規制情報のデータを用意した。

図 2 は、構築した Linked Data の概要、図 3 は、リソース間のリンクを示している。

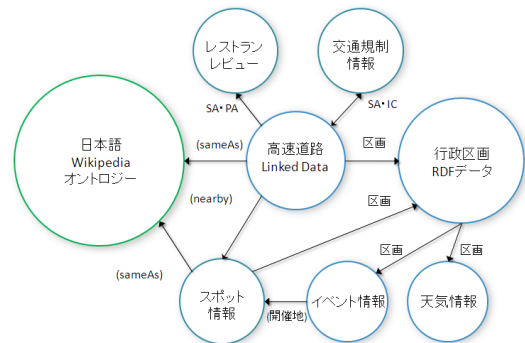


図 2 構築した Linked Data

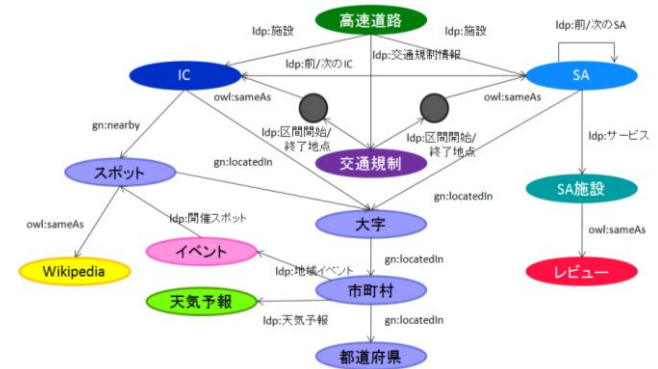


図 3 リソース間のリンク

データの更新性の度合いは、今回は、表 1 のように分類することにする。

表 1 データの更新性

更新性の低いデータ
高速道路 Linked Data (高速道路・SA・IC), スポット情報, 行政区画 RDF データ, レストランレビュー, 日本語 Wikipedia オントロジー
更新性の高いデータ
交通規制情報, 天気情報, イベント情報

外部の Linked Data として、日本語 Wikipedia オントロジー [玉川 10]の日本語 Wikipedia から抽出した RDF 形式のデータを利用している。

3.4 アプリケーション概要

開発したアプリケーションには、以下の4つの機能がある。

(1) ルート検索

出発地点・目的地のテキスト入力により、ルートを検索することができる機能である。詳細な経路のテキストと、地図上にポリラインを表示する。

ルート上の各ポイント(別の道路に入る、右折/左折する等)において、インターチェンジのポイントの場合は、Linked Data に対してクエリを投げ、インターチェンジの周辺情報、そのインターチェンジの位置する、市町村のイベント・天気情報へのリンクを表示させるためのマーカーを表示する。イベント情報のリンクをクリックすると、イベント情報の RDF データを取得し、開催地の緯度・経度情報から位置を特定し、その場所にマーカーを示し、イベントの概要を表示する、天気予報の場合は、地図の右側にデータを表示する。

地図上のマーカーを移動させ、現在地を指定することが可能である。交差点や高速道路の入り口等の、経路におけるマーカーも設置し、それをクリックすることでも現在地を指定できる。現在地のマーカーの緯度・経度に対して、ジオコーディングを行い、道路名を取得し、その道路名から、構築した Linked Data に対してクエリを投げ、それが高速道路でかつ、構築した Linked Data にデータが存在する場合は、サービスエリアの位置を取得し、地図上にマーカーとして表示する。

サービスエリアのマーカーから、リンクを貼り、サービスエリアのサービス一覧、さらに、レビューへのリンクを持っているレストランには、そのリンクを付与したことで、レビューを閲覧できる。また、サービスエリアからもインターチェンジ同様に、天気予報のデータも閲覧できる。

図 4 に、ルート検索の際のスクリーンショットを示す。



図 4 ルート検索のスクリーンショット

(2) スポット検索

道路名とキーワードを入力すると、その道路の周辺にあるキーワードに関連するスポットを検索することができる機能である。構築した Linked Data に、クエリを投げ、対象高速道路にあるインターチェンジの周辺にあるスポットのクラス情報に対して、入力されたキーワードで検索を実行している。日本語 Wikipedia オントロジーのインスタンスへのリンクを持っている場合は、リンクを辿り、対象インスタンスの持つ、情報(プロパティと値の対)を表示させることができる。さらに、対象インスタンスの属する、クラスも表示し、そのクラスに属するインスタンスの一覧も表示させることも可能である。対象スポットが周辺にあるインターチェンジとスポットの位置を地図に表示する。

(3) サービスエリア検索

道路名とキーワードと方向(上り/下り)入力とし、その道路にあるキーワードに関連するサービスエリアのサービス・施設等を検索することができる機能である。構築した Linked Data にクエリを投げ、対象高速道路にあるサービスエリアにある施設の名前とカテゴリに対して、入力されたキーワードで検索を実行している。「現在営業中の施設のみ検索」をチェックすると、現在時刻と曜日を取得し、現在営業中の施設のみに絞り込むことができる。営業時間は、曜日ごとに記述しており、SPARQL クエリを Endpoint に投げる際に、現在時刻が、その曜日の営業時間内であるかどうかの FILTER 句を追加して検索をかけている。

レビューデータへのリンクを持っているサービスエリア施設の場合は、レビューを表示させることができる。その施設のあるサービスエリアの位置を地図上に表示させる。

(4) 交通規制情報表示

道路名と方向(上り/下り)を入力とし、その対象道路の路線における最新の交通規制情報を一覧および、地図上に表示させる。交通規制情報の RDF と、データベースに格納された Linked Data にある対象高速道路の施設情報部分モデルのデータをメモリに用意し、検索をかけることで、規制開始/終了区間がサービスエリアやインターチェンジの場合、緯度・経度なども一度に合わせて取得することが可能になっている。

4. 検証・考察

4.1 日本語 Linked Data 構築における検証・考察

(1) リンクの生成方法の検証

レストラン等の店舗名は、表記ゆれ等が多く、データソースが異なる場合、店舗名だけで、同義であるインスタンスを発見し、データセット間のリンクを生成することは難しい。実際に構築した Linked Data においては、サービスエリア内のレストランから、レストランレビューへのリンクがその事例にあたる。提案手法に基づき、①サービスエリアの緯度・経度を中心として半径 300m 以内のレストランを検索②レストラン名の文字列部分照合の、位置の照合・名前の照合の 2 段階の照合をもとに、同一インスタンスを見つけ、リンクを生成した。その精度を検証する。

レストラン等の飲食サービス情報をもっている、サービスエリア(パーキングエリアを含む)全 44 カ所のデータをサンプルとして、リンク数と精度を調べた。図 5 がその結果である。

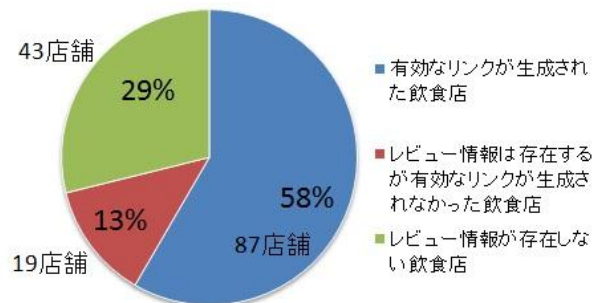


図 5 リンクの精度

サービスエリア全 44 カ所に、合計 149 の飲食店があり、106 店舗は、レビュー情報が存在した。その中の 87 店舗、約 82% は有効なリンクを生成できた。残りの 19 店舗は、レビュー情報は存在するが、有効なリンクを生成できていなかった。それらは、以下のような理由によるものであった。

- ・括弧書きなどの補足情報が、店舗名として含まれてしまっている
- ・別名が存在する
- ・名称が省略されている
- ・漢字表記と平仮名表記
- ・カタカナ表記とアルファベット表記
- ・サービスエリアの「フードコート」として何店舗かまとめられている

今回は、サービスエリアのデータと、飲食店のレビューデータのみの扱いであったが、レストランのマスターデータとしての Linked Data 用意し、それに、前述の日本語 Wikipedia オントロジーの場合と同様に、同義語・別称などの情報が、表記ゆれに対応するためには必要であると思われる。

(2) 課題や問題点

リンクの生成の他にも、データの正規化およびモデル化にあたって課題が多い。

本研究においては、各データに対し、モデル化を行い、RDF に変換していったが、店舗の営業時間や休業日等が複雑である場合等、データによっては、今回用意したモデルでは、十分に記述しきれないものもあった。例えば、店舗情報を記述するためのモデルなどといったように、各ドメインにおけるより汎用的なモデルを用意する必要があると思われる。

4.2 Linked Data を利用したアプリケーション開発における検証・考察

(1) アプリケーションの利用シナリオ

サービスの利用状況と対象ユーザは、

- ・走行中、助手席の同乗者による携帯等モバイル端末での利用
- ・家で、ドライブプラン計画立案のためにパソコン等での利用

を想定している。

開発したアプリケーションは以下のような利用シナリオが考えられる。

【シナリオ①】ルート検索

System: 出発地から目的地までのルートを表示
(目的地に向けて、東名高速道路に入った)

System: 東名高速道路にある SA を地図上に表示

User: 現在地の近くにある SA のサービスを知りたい

System: 現在営業中の店舗・サービスを表示

User: 現在営業中のレストランを発見。評判を知りたい。

System: 対象レストランのレビューを表示

User: 評価が良さそうなので、そのレストランのある SA に立ち寄る

【シナリオ②】スポット検索

(ドライブ旅行中、東名高速道路に入った)

User: 城を見に行きたい

System: 東名高速道路にある IC 付近の「城」に関連するスポット一覧表示

System: 日本語 Wikipedia オントロジーによる情報へのリンク提示

User: 城を3つ発見、位置や、日本語 Wikipedia オントロジーの情報である対象の城の城主や構造等を閲覧

User: 一覧の1つの城に立ち寄ってみることに決定

【シナリオ③】交通規制情報・天気情報

System: 東名高速道路(下り線)の現在の交通規制の一覧と地図を表示

System: 規制区間の手前の SA や IC も合わせて表示

User: 目的地の手前で雨による速度規制が続いていることを発見

User: その規制区間開始地点である SA の天気予報を知りたい

System: その SA の位置する市町村の天気予報を表示

User: 2時間後の予報は、警告もなく、天候も曇りであることを発見

User: 1つ手前の SA に立ち寄ってから目的地へ行くことに決定

(2) アプリケーション開発における課題・問題点

1点目に、SPARQL クエリの作成についてである。SPARQL のクエリを作成するには、クエリの対象となる RDF データの構造を熟知している必要がある。開発者が取得したい情報と SPARQL のクエリを結びつけるための機能があると便利である。そのためには、モデル化したデータの構造を詳細に記述するオントロジーが必要になると思われる。

2点目に、本研究では、比較的更新性の低いデータは、データベースに格納して利用した。しかし、サービスエリアの施設等は、決して不変なものではなく、レストランレビューデータは、増えていくものである。対象となる RDF グラフに対して、修正や削除・追加等が容易にできるようになることが、アプリケーションがユーザに、新しくかつ正しいデータを提供するために必要不可欠である。

さらに、本研究では、更新性に着目し、データベースに格納するデータと API 形式でデータを取得するデータとに分離したが、その分離の仕方をさらに検討し、処理時間とデータの分散の最適化を図る必要がある。

5. おわりに

本研究では、地理・位置データを中心とした日本語の Linked Data を構築し、それを利用したモビリティサービスを開発した。今回は、データの特性として更新性に着目し、それぞれに処理や利用方法を分けることで、Linked Data として様々なデータを扱うことのできる可能性について示し、情報サービスへの適用可能性について検証した。さらに、日本語の Linked Data 構築に向けての問題点や、情報サービス開発に必要なデータ検索を適切に実行できる RDF 記述のデータモデリング、どのようなクエリに対応することができるかについて考察した。従来の開発方法に比べ、容易に多様なクエリに対応できる可能性を示し、Linked Data の情報サービスへの適用可能性についても示すことができた。

参考文献

- [Becker 08] Christian Becker and Christian Bizer: “DBpedia Mobile: A Location-Enabled Linked Data Browser”, Web Semantics: Science, Services and Agents on the World Wide Web, vol.7, Issue4, pp.278-286, 2008
- [玉川 10] 玉川奨, 桜井慎弥, 手島拓也, 森田武史, 和泉憲明, 山口高平: 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, Vol.25, No. 5, pp.623-636, 2010