Linked Corporations Data in Japan

Shuya Abe, Yutaka Mitsuishi, Shinichiro Tago, Nobuyuki Igata, Seiji Okajima, Hiroaki Morikawa, and Fumihito Nishino

Fujitsu Laboratories Limited
{abe.shuya,mitsuishi-y,s-tago,igata,okajima.seiji,h.morikawa,nishino}@jp.fujitsu.com

Abstract. Based on the Open Data Charter of G8 in 2013, the Japanese government published a dataset covering approximately 4.4 million corporations in December 2015, but the dataset is merely rated as 3-star in the 5-star rating system of Open Data. To improve usability, we defined a schema for corporation data, converted the dataset into 5-star using this schema, and published the created dataset under Creative Commons Attribution 4.0 License. As far as we know, eight datasets currently refer to ours, which makes the degree of 5-star stronger. To demonstrate our dataset, we applied this dataset to a visualization system showing various information of a corporation.

Keywords: Linked Open Data, dataset, corporations, LOD4ALL

1 Introduction

The G8 leaders signed a charter on Open Data on June 18, 2013 [5]. The G8 members have also identified 14 high-value areas from which they will release data. One of them is the area of corporation register data.

Based on this background, the governments in the world aim to publish corporation datasets. Meanwhile, according to a survey of LOD datasets, no dataset of corporations is included in the highest referred datasets [8]. This survey indicates that the amount of corporation data is not sufficient in LOD datasets.

The objective of our work is to build a 5-star Japanese corporation register dataset. In Japan, the National Tax Agency (NTA) has published a dataset covering approximately 4.4 million corporations on December 01, 2015, but the dataset is merely rated as 3-star in the 5-star rating system. Our policy, which we believe is also common in the LOD community, is that low-star datasets must be converted into 5-star as early as possible for strengthening the power of LOD. Based on this policy, we defined a schema for corporation data, converted the Japanese dataset into RDF using this schema, appended links between our dataset and other data, and published the first version of this dataset under Creative Commons Attribution 4.0 License on December 09, 2015[1], only eight days after the publication date of the original dataset. As far as we know, eight datasets currently refer to ours, which makes the degree of 5-star stronger. We

[1] http://lod4all.net/datasetdetail.html?graph=http://lod4all.net/graph/corporate

also applied this enriched data to a visualization system for browsing a corporation.

The following sections describe our proposed schema design of the dataset, the building process and the demonstration. We finish the paper with future directions.

2 Building Linked Open Data

Tim Berners-Lee suggested a 5-star deployment scheme for Linked Data [2]. Since the approximately 4.4 million corporation register dataset published from NTA is in the XML format and is rated as 3-star, we have built an RDF-modeled dataset as 5-star and released the dataset.

First, we define the URIs of corporations that are independent of properties of a corporation and are not affected by changes in properties [7]. A URI is constructed with an identifier of a corporation and a code[2] of a country managing the identifier. The identifier must be independent of properties, unique among all the identifiers and persistent regardless of any changes happening to the corporation. Because the identifiers of NTA satisfy these features, we decided to employ the identifiers managing URIs. For example, the URI for the corporation with "1020001071491" as an identifier of NTA becomes "`http://lod4all.net/resource/corporate/jp/1020001071491`".

Next, we investigated the schema design [1,3], and we decided to adopt mainly "The Organization Ontology" [6] to represent a corporation in our schema design because it can describe properties of a corporation and various relationships between corporations.

We built an RDF-modeled dataset from the NTA dataset using the URIs and schema design. Additionally, we added a link to the corresponding URI in DBpedia[3] with `foaf:primaryTopic`, and attached the corresponding ticker symbol and EDINET code[4] as literals with `skos:notation` using a simple string matching method. We also added a geographical point of an address as literals with `geo:lat` and `geo:long` using a geocoding method.

The following example[5] shows a corporation in our dataset in Terse RDF Triple Language.

```
<http://lod4all.net/resource/corporate/jp/1020001071491> a org:Organization ;
    skos:prefLabel "Fujitsu Limited"@en ;
    org:identifier "1020001071491"^^<http://lod4all.net/ontology/corporate/jp/corporateNumber> ; # Corporation identifier of NTA
    skos:notation "6702"^^<http://lod4all.net/ontology/corporate/jp/stockTickerNumber> ; # Ticker symbol of Tokyo Stock Exchange
    skos:notation "E01766"^^<http://lod4all.net/ontology/corporate/jp/edinetNumber> ; # EDINET code
    foaf:primaryTopic <http://ja.dbpedia.org/resource/富士通> ; # Fujitsu
    dct:subject <http://lod4all.net/resource/corporate/jp/concept#株式会社> ; # Stock corporation
    dct:issued "2015-10-05T00:00:00"^^xsd:dateTime ;
    dct:modified "2015-10-15T00:00:00"^^xsd:dateTime ;
    org:hasRegisteredSite [
        a org:Site ;
        org:siteAddress [
            schema:postalCode "2110053" ;
            schema:addressCountry "Japan"@en ;
            gn:countryCode "JP" ;
            schema:addressRegion "Kanagawa"@en ;
```

[2] ISO 3166-1 alpha-2 country code: `http://www.iso.org/iso/country_codes`

[3] `http://dbpedia.org/`

[4] EDINET (`http://disclosure.edinet-fsa.go.jp/EKW0EZ1001.html`) is a system of Financial Services Agency of Japan for Annual Securities Report, and an EDINET code is an identifier of a corporation.

[5] Literals in the example are expressed in English with **@en** for understandability. They are actually expressed in Japanese with **@ja**.

```
        schema:addressLocality "Nakahara-ku, Kawasaki"@en ;
        schema:streetAddress "4-1-1 Kamikodanaka"@en ;
        skos:broadMatch <http://lod4all.net/resource/geo/神奈川県_川崎市_中原区> ; # Nakahara-ku, Kawasaki, Kanagawa
        geo:lat 35.58405 ; geo:long 139.6422 ;
    ] ;
] ;
org:changedBy [
    a org:ChangeEvent ;
    dct:subject 14a-corporate-jp-concept:吸収合併 ; # Absorption-type merger.
    prov:atTime "2015-10-15T00:00:00"^^xsd:dateTime ;
    dct:description
      "On October 1, 2015,
        Fujitsu Wireless Systems Limited (9030001085368) located at Nakasone 1376, Kumagaya, Saitama was merged."@en ;
] .
```

This example contains the data that are included in the dataset of NTA; a name (`skos:prefLabel`), an address (`org:hasRegisteredSite`), and an identifier (`org:identifier`). Additionally, information about important events such as M&A, and meta information about data update itself such as change of address (`org:changedBy`) is also included.

On December 09, 2015, we released the dataset on the SPARQL endpoint and the content negotiation framework in LOD4ALL[6] [4], an LOD utilization framework our group serves. We are updating it every week following NTA's daily update of the corporation register data.

To confirm utilization of our dataset by other people, we counted links from other datasets on LOD4ALL to ours. We found that as much as eight datasets currently link to our dataset.

3 Demonstration

We present a demonstration system for confirming the usefulness of our published dataset. The system inputs an entity from a user and displays an HTML page with multiple components, each of which is constructed with data linked to the entity.

Fig. 1 shows the snapshot for the corporation, Fujitsu Limited. If only the NTA-origin data were available, the snapshot would contain only the box denoted as NTA. By virtue of the enriched data we created, it contains various information as shown. For example, the DBpedia data obtained with links provides the logo in the upper left and the abstract in the upper right. Each component in the figure is generated by a JavaScript program with a function of querying a SPARQL endpoint. Below is the example of a SPARQL query used in the component showing the logo:

```
SELECT ?img WHERE {
    <http://lod4all.net/resource/corporate/jp/1020001071491> <http://xmlns.com/foaf/0.1/primaryTopic> ?dbpedia .
    ?dbpedia <http://dbpedia.org/ontology/thumbnail> ?img .
}
```

4 Conclusion

We defined a schema design of our Linked Data of corporations and published the dataset for about 4.4 million corporations by Creative Commons Attribution 4.0 License. Our dataset has already been linked from eight datasets, which indicates that our dataset contributes to enhancement of the LOD cloud. We also demonstrated our dataset using our visualization system that aggregates related data from various sources.

[6] `http://lod4all.net/`

Fig. 1. This snapshot consists of seven components with a title in a light blue header. For clarity, we grouped components and put them in red boxes with annotations. Each annotation means where the data comes from. The components display (1) the logo, (2) the location on the map based on `geo:lat` and `geo:long`, (3) the set of name, address and identifier, (4) the change information, (5) the EDINET code, (6) the abstract, and (7) group corporations.

Our future direction is enhancement of the value of our dataset. First, we will build additional LOD datasets of corporation register data from other countries. Secondly, we will append extra relationships to several other datasets. These expansions will contribute to the growth of data associated with corporations in LOD.

References

1. OpenCorporates, `http://opencorporates.com/`
2. Berners-Lee, T.: Linked Data (2006), `http://www.w3.org/DesignIssues/LinkedData.html`
3. Fukuyama, J.: The creation and trial offer of LOD in NDL: ISIL as target (2015), `http://current.ndl.go.jp/e1675`
4. Naseer, A., Kume, T., Izu, T., Igata, N.: LOD for All: Unlocking infinite opportunities. In: The Semantic Web Challenge 2014, The 13th International Semantic Web Conference (2014)
5. Office, U.C.: Open Data Charter (2013), `http://www.gov.uk/government/publications/open-data-charter`
6. Reynolds, D.: The Organization Ontology (2014), `http://www.w3.org/TR/vocab-org/`
7. Sauermann, L., Cyganiak, R.: Cool URIs for the semantic web (2008), `http://www.w3.org/TR/cooluris/`
8. Schmachtenberg, M., Bizer, C., Paulheim, H.: State of the LOD Cloud 2014 (2014), `http://linkeddatacatalog.dws.informatik.uni-mannheim.de/state/`