International Journal of Digital Information and Wireless Communications (IJDIWC) 6(3): 206-212
The Society of Digital Information and Wireless Communications, 2016 ISSN: 2225-658X (Online); ISSN 2412-6551 (Print)

KiwiLOD: A Framework to Transform New Zealand Open Data to Linked Open Data

Rivindu Perera and Parma Nand
Auckland University of Technology
Auckland 1010, New Zealand
{rperera, pnand}@aut.ac.nz

ABSTRACT

In this paper we present the theoretical framework to shift the New Zealand Open Data initiative to the next level by designing a scalable, queryable, and thus more accessible a Linked Data cloud, KiwiLOD. The KiwiLOD project will be executed in two stages. In the first stage a framework will be designed to transform the current open datasets (in comma separated and table formats) into Linked Data form. In the second stage, Natural Language Processing will be combined with an Information Extraction framework to transform the unstructured online and offline texts into Linked Data and made accessible via the cloud.

KEYWORDS

Open data, linked data, information extraction, Semantic Web, KiwiLOD

1 INTRODUCTION

The purpose of the open data initiative is to make publicly available the enormous amounts of non-personal data held by various institutions to enable better decision making in the modern economy. This initiative has been widely embraced by governments by publishing government data in open form, so that general public can access it for decision making. Furthermore, as published open data is in a structured form, it makes it easy for the users to easily analyze it and use it productively.

Open Government Data (OGD) [1] carries several benefits to both general public as well as for the government. OGD mainly help the public to get a clear and transparent view of the government process. For example, if the statistics related to annual expenses of a particular ministry is published as open data, this helps public to understand how that particular ministry manages its expenses. This is appreciated by many political scientists claiming that such approach can increase the trust in the general public. Using OGD, public can understand whether the government is performing well, and trace the achievements and targets that might not have been achieved as promised. On the other hand, by exposing the information to the public, the government can expect new ideas from the public. Furthermore, deep analysis of this government data can expose trends which can be valuable for decision makers (e.g., business managers) when planning for the long-term for strategies. In summary, OGD can increase the government transparency and public awareness of government processes.

Although governments have adhered and embraced the open data initiative, there are two major issues associated with the process. Firstly, the structured OGD is published in different formats and secondly the data encoded in unstructured text is not available as part of the structured open data.

The widely used formats to publish OGD are comma separated values (CSV), spreadsheets, HTML tables, and Portable Document Format (PDF) files. The issue that arises when publishing the data using multiple formats is that users face problems in linking multiple datasets and initiating a reasoning process. For example, assume that a user needs to find number of schools located in a particular area and the road traffic in that area in the last year. The user needs to open two data sets to complete for this analysis, and in addition he/she must do the analysis by manually comparing the data. It is obvious that this becomes more complex when performing analyses with multiple datasheets. Therefore, there is a clear need for infrastructure to link the raw data into a

domain classified entity framework. Much of the data in the web today are unstructured text [2]. This applies to government data as well, where a large quantity of data is stored as natural language text documents making it difficult for humans to understand quickly and making is almost impossible for computer systems to process the data.

The rest of the report is structured as follows. Section 2 provides a brief overview of New Zealand Open Data initiative. Section 3 offers an essential introduction to Linked Open Data and explains its basic functions. Section 4 is dedicated for open data to Linked Data transformation. In Section 5, we describes the Linked Data triple extraction from unstructured text. Section 6 lists some of the research outputs from this study. Section 7 describes future directions of this research while Section 8 concludes the report with a brief summary.

2 NEW ZEALAND OPEN DATA INITIATIVE

New Zealand government has also published open data following the OGD specification which belongs to multiple domains. This section provides an overview of the currently available New Zealand open data government datasets [3]. Table 1 shows properties of selected open government datasets.

Table 1 Sample set of Open Government Data and their properties

Category	Example datasets	Associated agencies	Formats
Agriculture, forestry, and fisheries	- Historical livestock production - Production, trade, and consumption of plywood	- Ministry for Primary Industries	XLS
Arts, culture, and heritage	- Survey of English Language Providers - New Zealand Heritage List	- Statistics New Zealand (SNZ) - Heritage New Zealand	XLS CSV
Building, construction, and housing	- Dwelling and Household Estimates - Market rent	- SNZ - Ministry of Business Innovation and Employment (MBIE)	CSV HTML
Commerce, trade, and industry	- Occupation Classification - Industrial Classification	- SNZ	XSL CSV
Education	- Directory of Educational Institutions - Teachers Register	- Ministry of Education	CSV DB
Employment	- Benefit Fact Sheets	- Ministry of Social Development	CSV
Energy	- Coal production statistics - Electricity Market Information - New Zealand Petroleum and Minerals GIS data services	- MBIE - Electricity Authority	XSL CSV
Environment and conservation	- Annual growing degree days - New Zealand Lizards Database	- Ministry for the Environment - Landcare Research	CSV DB
Fiscal, tax, and economics	- New Zealand Income Survey - Gross Domestic Product - Productivity Statistics - Institutional Sector Accounts statistics	- SNZ	CSV PDF HTML XSL
Health	- Tobacco Trends	- Ministry of Health	XSL
Infrastructure	- Internet Service Provider Survey	- SNZ	XSL

3 LINKED OPEN DATA

Linked Data is a set of guidelines to link related structured data on the web [4] [5]. The data is represented in triple form (⟨subject, predicate, object⟩) and. The subject and predicate in a triple are always IRIs (Internationalized Resource Identifiers). Specifying a triple as an IRI facilitates the linking with other triples.

Resource Description Framework (RDF) [6] is the basic representation format for the Semantic Web. RDF document is a directed graph where both nodes and edges are tagged with identifiers with two exceptions. Firstly, RDF allows for the encoding of data values as literals and secondly, blank nodes can be introduced in the RDF graph. Blank nodes are not labeled with IRIs and therefore these cannot be referred globally and to refer them in the context of the document a node identifier is provided. Figure 1 depicts an example RDF graph.

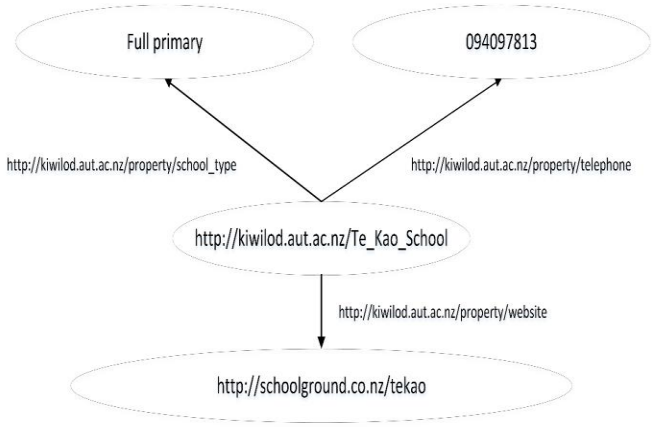

Figure 1. An example RDF graph

Ontology supports the organization of Linked Data based on the conceptualization. An ontology can be coded using both OWL (Web Ontology Language) and RDFS (RDF Schema). However, OWL is expressive than RDFS which is more suitable for small scale ontologies.

Technology stack for Linked Data is shown in Figure 2.

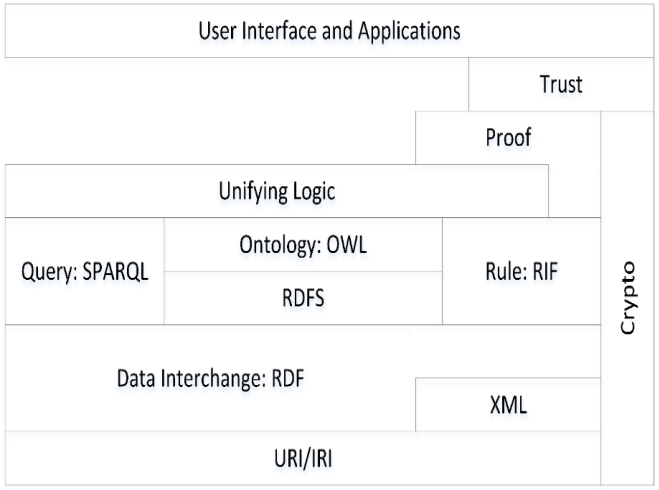

Figure 2. Technology stack for Linked Data

4 TRANSFORMING OPEN DATA TO LINKED OPEN DATA

This section describes the process of transforming Open Data to the Linked Open Data. The first step of the process is to build the conceptualization of the New Zealand entities, so that Linked Data triples can be correctly organized under classes. We have designed the version 1.0 of the ontology which is customized for New Zealand.

4.1 Ontology and Controlled Natural Language

The ontology contains 107 ontology classes (see Figure 3). Ontology can be further expanded and improved using the help of domain experts. However, a main drawback in managing an ontology with non-technical users is that they are not familiar with terminology of OWL and associated technical details. Therefore, to reach the non-technical domain experts, we provided the OWL in Controlled Natural Language (CNL) format. The CNL format converts the OWL specification of an ontology to a natural language representation using a limited terminology (shown in Figure 4). This can improve the non-technical community engagement in ontology expansion with an extended conceptualization of all of the different domains in New Zealand.

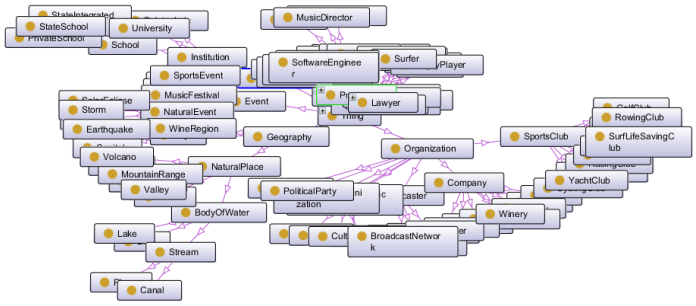

Figure 3 Ontology class view

Figure 4 Controlled natural language based description for ontology specification

4.2 Transforming Data to RDF/XML

New Zealand Open Data is published in four major document formats and all are in tabular format. In addition, many datasets share common specifications. We propose a template based approach to transformation taking aforementioned factors into consideration. The template is composed in XML and it contains necessary information to identify the triple elements and the data types associated with each object. A sample XML template developed to extract triples from "Schools Directory" CSV files is shown in Figure 5. These templates can be used with other datasets with minimal modifications.

```
<?xml version="1.0" encoding="UTF-8"?>
<template>
    <triple>
        <subject>col_1</subject>
        <predicate>colLabel</predicate>
        <object>value</object>
    </triple>
    <rules>
        <rule>noObjectValue=deleteTriple</rule>
        <rule>eliminate=col_1</rule>
    </rules>
    <datatypes>
        <datatype>school_number=string</datatype>
        <datatype>school_name=string</datatype>
        <datatype>telephone=string</datatype>
        <datatype>fax=string</datatype>
        <datatype>email=string</datatype>
        <datatype>principal=string</datatype>
        <datatype>school_website=string</datatype>
        <datatype>street=string</datatype>
        <datatype>suburb=string</datatype>
        <datatype>city=string</datatype>
        <datatype>postal_adress_1=string</datatype>
        <datatype>postal_adress_2=string</datatype>
        <datatype>postal_adress_3=string</datatype>
        <datatype>postal_code=string</datatype>
        <datatype>urban_area=string</datatype>
```

Figure 5 A sample XML template developed to extract triples

While extracting triples using the template, the process converts each subject to a URI. This is because RDF does not allow literal values for triple subject. The reason for this is that difficulty in linking data as well as ambiguity that may arise when two literal values exist with same content.

The framework has graphical user interface (GUI) to support RDF file creation and perform some basic edits (shown in Figure 6). This is because to provide opportunity to domain experts to decide the ontology class of the dataset. As current OGD is not classified under classes, it is not possible to assign an ontology class to a dataset without human involvement.

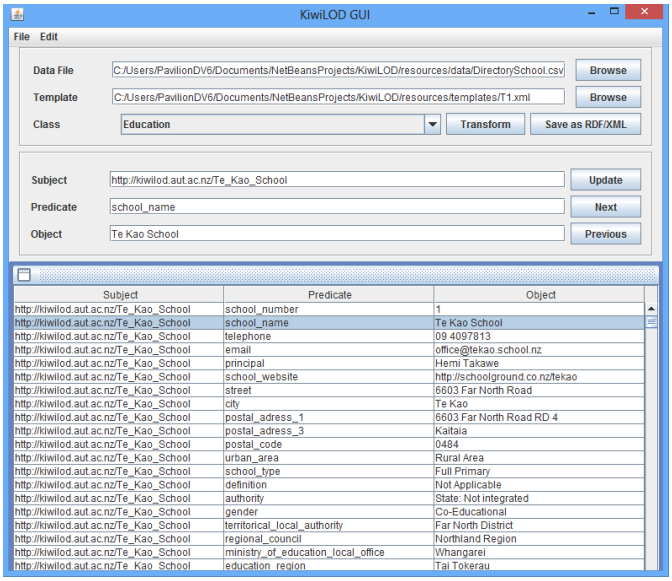

Figure 6 Graphical user interface of the framework

5 TEXT TO LINKED DATA

Although vast amount of data is available as structured open data, the majority of information is still provided as unstructured text. To build a rich Linked Dataset for New Zealand, there is a clear need to transform this unstructured data to Linked Data triples [7] [8] [9]. This section describes the framework developed to extract triples from unstructured text using Open Information Extraction (OpenIE).

In this phase we mainly focus on web based text resources related New Zealand, so that information embedded in such resources can be transformed to triples. These triples can enrich the Linked Data acquired from Open Data. However, we generalize the approach in way that we consider broad domain which belongs to the New Zealand without strictly limiting it to the basic government data. To accomplish such restriction, a pre-determined constraints are necessary which are not implemented as a part of this research.

5.1 Text Extraction and preprocessing

We use an automated web search to identify web pages which contains keywords extracted from a seed text resources. These seeds contains pre-determined text resources that we have already identified as which contain information directly related to New Zealand.

The keyword extraction is based on rule based noun phrase chunking method [8], [9]. We first Part-Of-Speech (POS) tag [10] the seed text resource and applied the predetermined rules to identify the keywords. The phrase chunking rules are shown in Table 2.

Table 2 Phrase chunking rules

Phrase chunking rules	
NN..	[JJ] [NN, NNP, NNS, NNPS]
NNP..	[JJR] [NN, NNP, NNS, NNPS]
NNS..	[NN, NNP, NNS, NNPS]
NNPS..	[NN, NNP, NNS, NNPS]

Once the text resources are identified, we extract the text using shallow text features. Basic set of these features include average word length, average sentence length, and the absolute number of words. In addition to these features, local context and heuristically derived set of features are also used. The local context is the position of the text in a document. Identifying this local context is essential in extracting text from HTML documents where text is placed in between boilerplate.

5.2. OpenIE

The traditional Information Extraction (IE), ClosedIE, focus on extracting relations using resource expensive linguistics models, hand tagged data, and hand crafted rules [2]. In essence, ClosedIE is targeting extracting relations which are specified in advance using rules. These models generally fails when applied in the heterogeneous text resources (i.e. web documents) [3] due to several reasons. Rules generated to one domain sometimes conflict with a different domain. Next, tagging must be carried out in all domains. However, in a large-scale heterogeneous dataset identifying all domains makes a challenge itself. Finally, if text resource is added from a completely new domain, that makes the ClosedIE a complete failure.

To address this incapability of traditional IE models, Open Information Extraction (OpenIE) [11] [12] [13] is introduced. Open IE utilizes the self-supervised learning where relation extraction process is deliberated as a complete automatic process eliminating deep linguistic parsing.

According to Wu and Weld [14], OpenIE process is a function from a document (d), to a set of triples $\langle arg_1, rel, arg_2 \rangle$ where the *args* are noun phrases and *rel* is a textual fragment which indicates an implicit, semantic relation between the two *args*. An important feature of OpenIE is that one triple is generated for every relation appeared in the text. However, recent progress in the OpenIE has moved beyond this by generating new triples using existing ones though template and seed based mechanisms.

When considering a number of document (D), the complexity of OpenIE is O(D). For the same

documents collection and set of relations (R) which are specified in advance, ClosedIE has the complexity of O(D*R). In particular, we utilize a clause based OpenIE model [11] to extract relations. The clauses include 7 basic clause types and 5 other extended clause types.

These clauses are identified by dependency parsing a seed sentence set. Each dependency pattern extracted from parse tree is mapped to a clause type. However, the available dependency parse is not compatible with recent Stanford dependency parse specifications which include modified set of dependency parse where some are aggregated while some are extended. The parser used for this research includes the latest dependency parse specification based clause extractor which is modified accordingly. These dependencies are then converted to clauses and propositions.

After applying the OpenIE, the resulting relations are converted to triple form. The arg_1 is mapped as the subject with a URI, *rel* as the predicate, and the arg_2 is mapped as the object. A key factor to consider here is that except the subject, it is difficult to automatically identify objects with URIs. This is because object values cannot be classified as literals and as real world conceptualized entities. For this we propose the crowd-sourced validation.

Figure 7 illustrates the number of triples extracted from the sentence collection. According to the figure, it is clear that framework has extracted reasonable number of triples from the sentences.

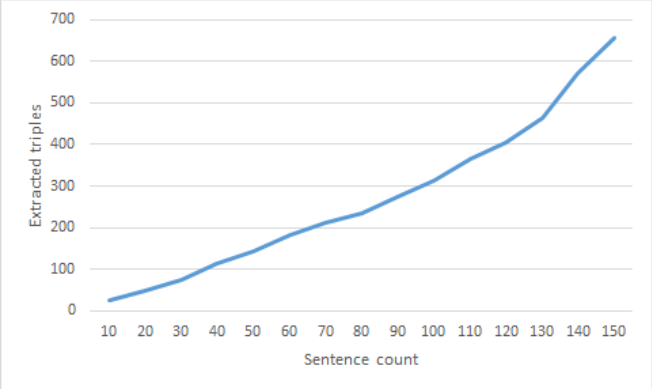

Figure 7 Extracted triples for sentence collection

5.3. Triple Storage

The extracted triples can be stored in a graph database to build a scalable Linked Data cloud and open it to the general public. As an additional task we have imported the RDF triples to GraphDB community edition based graph database.

6 CONCLUSION AND FUTURE DIRECTIONS

The aim of this research was to examine the possibility of developing a Linked Data cloud based on the New Zealand OGD. Initial steps of this research designed an ontology for the New Zealand OGD and a process of transforming existing data to triples using a template based approach. The later steps of this project investigated the methodology of extracting triples from unstructured data to enhance the triple storage.

The research here uses limited set of data files and templates to carry out an investigation of transforming OGD to Linked Data. These experiments and implementations shows that the method is viable. Therefore, as future work of this research one should focus on building templates and transforming existing data to Linked Data while new data can be stored in the graph database as triples. Another key factor to consider here is representation of statistical data in the Linked Data cloud. Statistical data slightly differs from other data as they contain archived records of annually collected data. These can be transformed in the same form as the other open data. However, there are alternative ways of organizing this data based on time, category, or based on other provided factors. The current research did not consider such classifications hence is left as future work.

The process of triple extraction from unstructured text includes a text extraction and preprocessing module in which boilerplate of web based text is removed and co-references are resolved. However, with the meta web [15] concept, currently there is a growing trend towards populating web pages with meta data. The future research focus of improving this module to extract metadata from

webpages to associate with extracted text. This metadata can help later in the process to classify the triples to ontology classes and enrich the data with tags.

Triple extraction is based on OpenIE model which incorporates clauses to extract large number of relations from the text. However, the current model has some limitations. Among them the main drawback is that the relation extraction process does not support information which is represented with different forms. For example, Wikipedia shows the birthdate of a person in the form of "Steve Jobs (18-09-1949)". These phrases do not contain clauses. Therefore, future research will focus on extracting such relations with rule based and trained classifiers.

The current research utilizes GraphDB for triple storage. However, there are several triple storages which can be utilized for this task. Table shows a comparison of some of the triple storages which we can utilize for this task. The future research will provide a practical view of using these triple storages with multiple experiments carried using OGD. Furthermore, we expect to build a web service to automate the process explained in this research and connecting the triple storage. The web service will provide the same set of functionalities using an easy to use interface.

REFERENCES

1. Zapilko, B., Mathiak, B. Performing statistical methods on linked data. In: International Conference on Dublin Core and Metadata Applications. Copenhagen, Denmark: European Library Office (2011).
2. Sirmakessis, S. Text Mining and Its Applications: Results of the NEMIS Launch Conference. 138th ed. Springer International Publishing (2012).
3. Oh, J. Open Data in New Zealand. University of Auckland. Auckland, New Zealand (2013).
4. Gerber, D., Ngomo, A. Bootstrapping the Linked Data Web. In: The 10th International Semantic Web Conference. Bonn: Springer-Verlag (2011).
5. Ngomo, A., Auer, S., Lehmann, J., Zaveri, A. Introduction to Linked Data and Its Lifecycle on the Web. In: 7th International Summer School. Galway, Ireland: Springer International Publishing (2014).
6. Segaran, T., Evans, C., Taylor, J. Programming the Semantic Web. vol 54 (2009).
7. Perera, R., Nand, P. A Multi-strategy Approach for Lexicalizing Linked Open Data. In: 16th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing). Springer International Publishing (2015).
8. Perera, R, Nand, P. The Role of Linked Data in Content Selection. Trends Artif Intell. (2014).
9. Perera, R, Nand, P. RealText cs - Corpus Based Domain Independent Content Selection Model. In: 26th IEEE International Conference on Tools with Artificial Intelligence. IEEE Press (2014).
10. Manning, C., Bauer, J., Finkel, J., Bethard, S., McClosky, D. The Stanford CoreNLP Natural Language Processing Toolkit. In: The 52nd Annual Meeting of the Association for Computational Linguistics. Baltimore: Association for Computational Linguistics (2014).
11. Corro, L., Gemulla, R. ClausIE: clause-based open information extraction. (2013).
12. Mausam, S., Bart, R., Soderland, S., Etzioni, O. Open language learning for information extraction. In: Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Jeju Island: Association for Computational Linguistics (2012).
13. Zhila, A., Gelbukh, A. Comparison of open information extraction for Spanish and English. In: International Dialogue Conference. Moscow, Russia: Association for Computational Linguistics (2013).
14. Wu, F., Weld, D. Open Information Extraction using Wikipedia. In: 48th Annual Meeting of the Association for Computational Linguistics. Uppsala, Sweden: Association for Computational Linguistics (2010).
15. Ceri, S., Bozzon, A. Web Information Retrieval. Springer International Publishing (2013).