Converting Governmental Datasets into Linked Data

Timothy Lebo
lebot@rpi.edu

Gregory Todd Williams
willig4@cs.rpi.edu

Tetherless World Constellation
Rensselaer Polytechnic Institute
110 8^{th} Street, Troy, NY 12180

ABSTRACT

Linked Data provide many benefits to data consumers, but many publicly available datasets are still released in the Comma Separated Values (CSV) format, a ubiquitous common denominator. We introduce a methodology to transform such datasets into Linked Data. Our design is based on requirements identified while surveying existing governmental datasets released by data.gov. We present an implementation-independent RDF vocabulary to describe how a CSV dataset should be promoted into Linked Data, and use a Java-based converter to produce 5.3 billion RDF triples from 312 data.gov datasets.

Categories and Subject Descriptors

E.2 [**Data Storage Representations**]: Object representation

General Terms

Linked Data

Keywords

Government data, interpretation, incremental enhancement

1. INTRODUCTION

Although RDF and Linked Data provide many benefits, other factors influence an organization's choice of data representation. When organizations publish their data in alternate forms, its consumers must choose to adopt potentially ad hoc conventions or disregard the data's potential. The Comma Separated Values (CSV) format remains a ubiquitous common denominator for the transfer of tabular data, likely because of its simple text-based structure and ease of processing with a wide variety of tools. Unfortunately, the simple structure of the CSV format makes interpretation of the data difficult without domain knowledge, accompanying "codebook" data, and manual data inspection.

I-SEMANTICS 2010, September 1-3, 2010, Graz, Austria.

We present an approach which supports motivated parties to produce quality Linked Data from third-party CSV sources. Because deciphering others' data can be time consuming, we strive to minimize the initial time commitment required to produce RDF. Using only three parameters that describe the dataset, RDF can be automatically produced in a structure amenable to further enhancement and linking. As time, motivation, and domain understanding permit, incremental enhancements can improve the data quality, structural modeling, or data linking.

2. DATA CONVERSION

Our approach provides stable RDF data that can be improved without disrupting existing applications that use it. Importantly, each incremental enhancement produces a new dataset combining existing RDF with newly enhanced data. Data from previous enhancements are never changed, only added to. In this way, applications may rely on the types and structure of data available at the time they were written without concern that said data will be changed by further enhancements. A full description of the URI scheme that enables this stability is beyond the scope of this paper[1]

In converting CSV data to Linked Data, we distinguish between two types of conversion processes: raw conversions and enhancement conversions. Raw conversions are used to convert CSV data to simple RDF while enhancement conversions are used to iteratively increase the quality of the RDF by casting values to datatypes and resources, restructuring relationships, and linking to external ontologies and datasets.

Both types of conversion process are driven by declarative parameters encoded using RDF. This RDF not only prescribes how conversion tools should process the input data, but also provides a provenance history of the operations that led from the original CSV source to published Linked Data. The conversion ontology used for encoding the declarative parameters is identified by the URI `http://purl.org/twc/vocab/conversion/` and is abbreviated by the prefix `conv:` in the following discussion. Due to space limitations in this paper, URIs segments of the form `/source/data-gov/dataset/NNN/version/` are abbreviated using the form `/s/data-gov/d/NNN/v/`.

2.1 Raw Conversion

Raw conversion is designed to make CSV data quickly accessible to existing RDF tools. Although the conversion

[1]Additional materials are available at http://data-gov.tw.rpi.edu/wiki/Triplify_challenge_2010

is quick, many characteristics that Linked Data consumers expect and desire are left for subsequent enhancement conversions (see Section 2.2). To illustrate this, two sample rows based on data.gov dataset 1450 are shown in Table 1, while the triples resulting from both raw and enhancement conversions are shown in Figure 1. To convert a CSV table into an RDF graph, a URI is minted for each row and described with predicates derived from each column. The column's cell value becomes the triple's plain literal object. Raw conversion may be parameterized to handle structural variances such as header rows that appear after the first row (due to titles and captions) and data rows that finish before the last.

2.2 Enhancement Conversions

Enhancement conversion takes an existing RDF dataset as input and produces new RDF triples by applying enhancement operations to the input data. The parameter ontology uses the terms domain, range, subproperty, and subclass in ways that reflect the RDFS semantics. Below we describe the primary enhancement operations, the parameters that invoke them, and their resulting RDF. A full example of enhancement parameters is shown in Figure 1.

2.2.1 Row Typing

Because URIs created for CSV rows are central to the resulting RDF, it should be easy to understand and select them as members of a domain-relevant class. For example, rows in Dataset 1491 represent Disasters, while rows in Dataset 1492 represent Disaster Aid Obligations. These can be typed using `conv:domain_name` on any property enhancement.

2.2.2 Casting to Datatypes and Resources

Casting turns plain literals into URIs or typed literals including XSD numerics, xsd:boolean, xsd:date and xsd:dateTime. Parameters can be added to handle untraditional interpretations. For example, as shown in Table 1, a "*" represents true and an empty value represents false. Pairs of `conv:symbol` and `conv:interpretation` can be used as shown in Figure 1. Similarly, date patterns such as "M/d/yy" may be provided to allow parsing a wide range of `xsd:date`s and `xsd:dateTime`s.

2.2.3 Resource Promotions

Values may be promoted to resources in several more complex ways than simple casting. When this promotion occurs, the raw literal being promoted becomes an `rdfs:label` of the promoted resource. *Property-scoped promotion* converts all the values originating from a specific CSV column into resources in a dataset- and column-specific value space. For example, Dataset 1564 cites a "Marketing Category" for animal-tested drugs with values "NADA", "ANADA", and "UNAPPROVED OTHER" and would be promoted to `/s/data-gov/d/1564/value-of/marketing_category/NADA` *Typed promotion* is similar to property-scoped promotion with the exception that the resource value space is specified by a type. If two columns contain state abbreviations, both columns would be promoted with the type "state" and result in identical URIs. For example, the columns in Dataset 1147 indicating Alabama's FIPS code of "01" as an "origin" or "destination" state of migration flow are promoted to the form `/s/data-gov/d/1147/typed/state/01`.

Column bundling introduces a new resource that takes over "ownership" of a number of existing property-value pairs for each row instance. For example, the columns in Dataset 1171 containing the prefix, first name, middle name, last name, and suffix of members of Federal Advisory Committee Act committees are associated with a URI representing the person instead of the row of the CSV.

Finally, *crutch promotion* allows a promoted value's URI to be constructed based on several property values. URI construction is based on a template parameter that specifies the relevant property values. For example, the "District" column in Dataset 1330, which references congressional districts by number, can be crutch-promoted using the value of the "state" column using the pattern "[@state]-[@district]", resulting in a URI of `/s/data-gov/d/1330/value-of/district/CA-1`. In this sense, "District" uses "State" as a "crutch" to be promoted to a resource that stands on its own.

2.2.4 Subclass and Subproperty Linking

Linking to external ontologies may be performed at the class or property level. In each case, the local name of the class or property is cited along with the external property or class. Each subproperty or subclass enhancement results in an additional triple for each involved resource. For example, Figure 1 describes the "legal_entity_name" property as a subproperty of `foaf:name`, resulting in two `foaf:name` triples.

2.2.5 Object sameAs Linking

Promoted resources can be linked to external resources with `owl:sameAs`. Literal values used to create the resources are matched with values in a linking file containing descriptions of external resources. For example, in Dataset 1492, the state value "Texas" is matched with a triple in the linking file `dbpedia:Texas dc:identifier "Texas"`, yielding `</s/data-gov/d/1492/v/2010-Jan-21/typed/state/Texas> owl:sameAs dbpedia:Texas`. Multiple linking files may be specified at the same time to expand the scope of external linking, and resources may have any number of identifiers. Linking files can be reused and augmented for subsequent dataset linking.

2.3 Application to Data.gov Datasets

Data.gov's simple and consistent URL structure allows for straightforward automation to retrieve available data files. We retrieved all CSV datasets available from data.gov and converted them to RDF using the approach described above. Of the 1780 datasets listed in the Data.gov Raw Data Catalog, 312 datasets are available as CSV (some datasets providing more than one CSV file). In all, we retrieved 992 CSV files and converted them into 5.3B triples using the raw conversion parameters. The process of converting all 992 CSV files to RDF took under 12 hours on a Dell PowerEdge T610 with dual Xeon E5504 quad-core 2.0GHz CPUs.

Because enhancements require a degree of domain understanding, the raw conversions have been supplemented with only 24 million triples to date. These 24 million triples were the result of one of the authors spending 2 days investigating datasets and producing enhancement parameters; we believe enhancing other datasets would yield similar results, even with a minimally-trained user identifying the enhancement parameters. So far, the rows in 4

Table 1: Example data based on data.gov dataset 1450.

State	Offers Plans In This State Only	Legal Entity Name	Organization Name
Alabama		ACCENDO INSURANCE COMPANY	RxAmerica
Florida	*	SUMMACARE INC.	SummaCare

```
# ---------------------------------------- CONVERSION PARAMETERS ----------------------------------------
_:dataset a void:Dataset; conv:base_uri "http://data-gov.tw.rpi.edu"^^xsd:anyURI;
    conv:source_identifier "data-gov"; conv:dataset_identifier "1450"; conv:dataset_version "18-May-2009";
    conv:conversion_process [
        conv:enhancement_identifier "1";
        conv:enhance [ conv:property_name "state"; conv:range rdfs:Resource;
                       conv:range_name "State";
                       conv:links_via <http://rpi.edu/~lebot/lod-links/state-fips-geonames.ttl>;
                       conv:subject_of dc:identifier ];
        conv:enhance [ conv:property_name "offers_plans_in_this_state_only"; conv:range xsd:boolean;
                       conv:interpret [ conv:symbol "*"; conv:interpretation true ];
                       conv:interpret [ conv:symbol "" ; conv:interpretation false ] ];
        conv:enhance [ conv:property_name "legal_entity_name"; conv:bundled_by _:org_bundle;
                       conv:subproperty_of foaf:name ];
        conv:enhance [ conv:property_name "organization_name"; conv:bundled_by _:org_bundle ] ].
_:org_bundle a conv:ImplicitBundle; conv:type_name "Organization"; conv:property_name "organization".
# ---------------------------------------- RESULTING TRIPLES ----------------------------------------
ds1450:thing_2 raw:state "Alabama" ; raw:offers_plans_in_this_state_only "" ;
    raw:legal_entity_name "ACCENDO INSURANCE COMPANY" ; raw:organization_name "RxAmerica" ;
    e1:state state:Alabama ; e1:offers_plans_in_this_state_only false ;
    e1:organization org:organization_1; conv:csvRow 2 .
state:Alabama a ds1450_vocab:State ; rdfs:label "Alabama" ; owl:sameAs <http://sws.geonames.org/4829764/> .
org:organization_1 a ds1450_vocab:Organization ; e1:organization_name "RxAmerica" ;
    e1:legal_entity_name "ACCENDO INSURANCE COMPANY" ; foaf:name "ACCENDO INSURANCE COMPANY" .

ds1450:thing_3 raw:state "Florida" ; raw:offers_plans_in_this_state_only "*" ;
    raw:legal_entity_name "SUMMACARE INC." ; raw:organization_name "SummaCare" ;
    e1:state state:Florida ; e1:offers_plans_in_this_state_only true ;
    e1:organization org:organization_2; conv:csvRow 3 .
state:Florida a ds1450_vocab:State ; rdfs:label "Florida" ; owl:sameAs <http://sws.geonames.org/4155751/> .
org:organization_2 a ds1450_vocab:Organization ; e1:organization_name "SummaCare" ;
    e1:legal_entity_name "SUMMACARE INC." ; foaf:name "SUMMACARE INC." .
```

Figure 1: Enhancement parameters and resulting triples for the dataset shown in Table 1.

datasets are typed to domain-specific classes. 1,546 properties are datatype properties with the following ranges: xsd:gYear (6), xsd:boolean (7), xsd:date (12), xsd:dateTime (12), xsd:nonNegativeInteger (78), xsd:integer (306), and xsd:decimal (1,125). Seven properties are bundled. Of 120 properties promoted to resources, 30 are typed, 4 use crutches, and 12 are linked with `owl:sameAs`. Of the 30 typed promotions, 4 subclass FOAF or W3C's WGS84 classes. Finally, the 12 linked properties lead to 950 resources asserted as `owl:sameAs` to resources in DBpedia, GovTrack.us, and GeoNames.

3. CONCLUSIONS

We have presented an approach to convert third party CSV data into Linked Data in a way that allows incremental enhancements to stably augment its initial conversion. Applying this approach, we have converted a large amount of US governmental data into RDF and have begun the process of enhancing it with data typing, restructured relationships, and linking into the Linked Open Data cloud. Based on this experience we believe the enhancement of data.gov data can progress quickly and with minimal effort by a broader community. We hope that this methodology can be used to bootstrap similar Linked Data publishing efforts involving existing CSV data sources, and help inspire data providers to adopt semantic web techniques in their future data publication strategies.

We see two primary challenges to applying the capabilities we present more broadly. First, an appropriate user interface is required to increase the accessibility of these capabilities. Second, a creative solution to enable and motivate distributed, ad hoc communities to join forces to promote and link CSV datasets would provide impetus for continual contributions. The approach we present here would underlie this ecosystem and may require additional development to support these additional requirements.

4. ACKNOWLEDGMENTS

We thank Alvaro Graves Fuenzalida and Jesse Weaver for their help in developing this submission.