# A Lightweight Approach to Explore, Enrich and Use Data with a Geospatial Dimension with Semantic Web Technologies

Christophe Debruyne
Trinity College Dublin, Dublin 2, Ireland
first.last@adaptcentre.ie

Kris McGlinn
Trinity College Dublin, Dublin 2, Ireland
first.last@adaptcentre.ie

Lorraine McNerney
Ordnance Survey Ireland, Dublin 8, Ireland
first.last@osi.ie

Declan O'Sullivan
Trinity College Dublin, Dublin 2, Ireland
first.last@adaptcentre.ie

## ABSTRACT

The concept of "location" provides one a useful dimension to explore, align, combine, and analyze data. Though one can rely on bespoke GIS systems to conduct their data analyses, we aim to investigate the feasibility of using Semantic Web technologies to leverage the exploration and enrichment of data in CSV files with the vast amount of geographic and geospatial data that are available on the Linked Data Web. In this paper, we propose a lightweight method and set of tools for: uplift – transforming non-RDF resources into RDF documents; creating links between RDF datasets; client-side processing of geospatial functions; and downlift – transforming (enriched) RDF documents back into a non-RDF format. With this approach, people who wish to avail of the spatial dimension in data can do so from their client (e.g., in a browser) without the need to rely on bespoke technology. This could be of great utility for decision makers and scholars, amongst others. We applied our approach on datasets that are hosted on the Irish open data portal, and combined it with authoritative geospatial data made available by Ordnance Survey Ireland (OSi). Albeit aware that our approach cannot compete with specialist tools, we do demonstrate its feasibility. Though currently conducted for enriching datasets hosted on the Irish open data portal, future work will look into broader governance and provenance aspects of geospatial data enriched dataset management.

## CCS CONCEPTS

**•Information systems → Resource Description Framework (RDF); Geographic information systems;**

## KEYWORDS

GeoSPARQL, Linked Data, Interlinking, Data enrichment, Ordnance Survey Ireland

**ACM Reference format:**
Christophe Debruyne, Kris McGlinn, Lorraine McNerney, and Declan O'Sullivan. 2017. A Lightweight Approach to Explore, Enrich and Use Data with a Geospatial Dimension with Semantic Web Technologies. In *Proceedings of GeoRich'17 , Chicago, IL, USA, May 14, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3080546.3080548

## 1 INTRODUCTION

Linked Data [2] is a set of best practices and guidelines to publish and interlink data on the Web by cleverly combining several standardized technologies: the Resource Description Framework[1] (RDF) to describe things; URIs to identify those things – and in particular HTTP URIs to retrieve these descriptions over the Web; and content-negotiation to retrieve the desired format (HTML for users, RDF for machines, etc.). The importance of geospatial data can be observed by the vast amount of geographic or geospatial datasets available on the Linked Data Web. And as most datasets also have a geospatial dimension that is either explicit or implicit, it comes as no surprise that "location" is a convenient way for aligning and combining different datasets [14]. Datasets can also rely on standardized vocabularies such as GeoSPARQL[2], for representing and querying geospatial information.

One can rely on bespoke systems to analyze data with a geospatial dimension; e.g., ArcGIS[3] for desktop analysis or GeoSPARQL-enabled triplestores such as Parliament [1] to avail of spatial predicates in RDF. We believe, however, that engaging with geospatial data without relying upon bespoke approaches can be enabled by means of an adequate set of best practices and tools. In this paper, we present a lightweight method and set of tools for two typically desired processes: *uplift* – transforming non-RDF resources into RDF documents; creating links between RDF datasets; client-side processing of geospatial functions; and *downlift* – transforming (enriched) RDF documents back into a non-RDF format. We furthermore demonstrate the feasibility of client-side processing of geospatial queries in GeoSPARQL to leverage the use of geospatial data on the Linked Data web and to facilitate the aforementioned method.

The remainder of this paper is organized as follows: Section 2 provides background information on Linked Data, GeoSPARQL, query technologies, and the context in which this study was undertaken; Section 3 presents our method from the transformation of non-RDF resources into RDF and combining the resulting RDF with other, external datasets, to transforming it back in its original format; Section 3 furthermore outlines the algorithms used in the various tools; Section 4 illustrates our approach using a simple open dataset from `data.gov.ie`; we discuss our results in Section 5 prior to concluding and outlining future work in Section 6.

---

*GeoRich'17 , Chicago, IL, USA*

DOI: http://dx.doi.org/10.1145/3080546.3080548

---

[1] https://www.w3.org/RDF/

[2] http://www.opengeospatial.org/standards/geosparql

[3] http://desktop.arcgis.com/

## 2 BACKGROUND

In this section, we briefly elaborate on some of the concepts that will be used throughout this paper. We assume the reader is familiar with Linked Data, the Resource Description Framework (RDF), the Terse RDF Triple Language[4] (or TURTLE) RDF serialization format, and SPARQL[5] for querying RDF graphs.

### 2.1 GeoSPARQL

The OGC GeoSPARQL standard proposes two things. First, it provides a vocabulary to represent geographical features and geometries, of which the latter can be represented in either Well-Know Text (WKT) format or Geography Markup Language (GML). Features can be related with predicates such as `geo:hasGeometry`, or specializations thereof. Secondly, as its name implies, it defines an extension to SPARQL for formulating geospatial queries using predicates such as `geof:sfDisjoint`, `geof:sfIntersects`, etc. Assuming a GeoSPARQL-enabled endpoint with descriptions of counties, the query in Listing 1 would return pairs of labels, in English, of counties in Ireland that are disjoint.

**Listing 1: GeoSPARQL query for returning pairs of labels in English of counties that are disjoint.**

```
PREFIX osi: <http://ontologies.geohive.ie/osi#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#">
PREFIX gf: <http://www.opengis.net/def/function/geosparql/>
SELECT ?c1l ?c2l {
 ?c1 a osi:County ; rdfs:label ?c1l ; geo:hasGeometry ?g1 .
 ?c2 a osi:County ; rdfs:label ?c2l ; geo:hasGeometry ?g2 .
 FILTER (?c1 != ?c2)
 FILTER langMatches( lang(?c1l), "en" )
 FILTER langMatches( lang(?c2l), "en" )
 ?g1 geo:asWKT ?w1 .
 ?g2 geo:asWKT ?w2 .
 FILTER(gf:sfDisjoint(?w1, ?w2))
}
```

### 2.2 Triple Pattern Fragments

Scholars realized that there are two extremes observable on the Semantic Web [17]; RDF dumps and SPARQL endpoints. RDF dumps do not allow agents to interrogate the data, but limits the resources required by the server. SPARQL endpoints, on the other hand, allow agents to formulate queries, but requires more resources on the server. A Linked Data front end does allow an agent to obtain an RDF description to be dereferenced from an HTTP URI, yet does not allow clients to interrogate the data. In [17], the authors proposed Triple Pattern Fragments (TPF), which provides a compromise by breaking down queries into simple queries (based on triple patterns) that the server needs to return and the client uses these to compute the result set. The load is thus distributed between server and client. TPF, as of yet, provides no support for client-side processing of GeoSPARQL functions. In [3], we proposed to extend a TPF client with GeoSPARQL functions. We will use this extension as part of our tool suite for engaging with geospatial-enriched data on the Linked Data Web.

## 3 APPROACH

Fig. 1 depicts an overview of the processes and artifacts involved in our method. We will, for each of the processes, describe our approach and the tools used.

### 3.1 Uplift of CSV Data into an RDF Dataset

The generation of RDF from non-RDF documents is called *uplift*. This is achieved via a mapping. There are two types of mappings available; a direct mapping – which reflects the structure of the document – and a declarative mapping allowing one to relate structure of the document to vocabularies. One can see that the former results in RDF documents that are only meaningful to those who understand what is contained in a CSV file, and how it is structured. The advantage of the former is that it does not rely on the creation of a mapping; the mapping is implemented in an algorithm.

Our approach lies somewhere in the middle. Given a CSV file, we generate a declarative mapping based on R2RML[6] – a W3C standard for mapping relational databases to RDF by treating the CSV file as one table and column headers as fields. In the generated mapping, we also keep track of information about the CSV file's structure – e.g., the original column names as they might not be SQL compliant, the order of the columns, and consider each line to be an instance of a concept called "Record". These three additions will later help us to conduct a *downlift* of the enriched RDF.

The approach of generating a mapping also allows one to add additional mappings to facilitate subsequent processes, as R2RML allows for a certain degree of manipulation of values in the mapping. For example, in an Irish context, coordinates are often captured in Irish Transverse Mercator (ITM). Unlike relational databases with geospatial support, one cannot rely on underlying technology to transform the coordinates into WGS84. R2RML does not provide support for functions, but in prior work we proposed an extension to R2RML, called R2RML-F [4], in which these functions can be expressed in JavaScript, *and* are part of the mapping. In fact, the transformation of ITM into WGS84 was the demonstrator in [4].[7]

The result of executing the R2RML-F mapping is an RDF file represent in triples the data from the CSV. Note that users can already engage with this RDF file by using SPARQL, as discussed later in this paper.

### 3.2 Creation of Links with other Datasets

The creation of links with other datasets can be achieved in various ways that depend on the nature of the dataset, the nature of the Linked Data dataset to be linked with, and the requirements of the linking exercise. In this section, we will cover and discuss some ways to create links. The decision as to which to choose, however, is deemed outside the scope of this paper.

We will cover two broad ways to create links with other Linked Datasets: the application of link discovery tools and the use of SPARQL CONSTRUCT queries. First, quite a few link discovery tools exist. Silk [8], for instance, allows one to declare how RDF

---

[4] https://www.w3.org/TR/turtle/

[5] https://www.w3.org/TR/sparql11-query/

[6] https://www.w3.org/TR/r2rml/

[7] The old prototype was based on a limited implementation of R2RML. A newer prototype, with support for named graphs and functions is available at https://opengogs.adaptcentre.ie/debruync/r2rml.

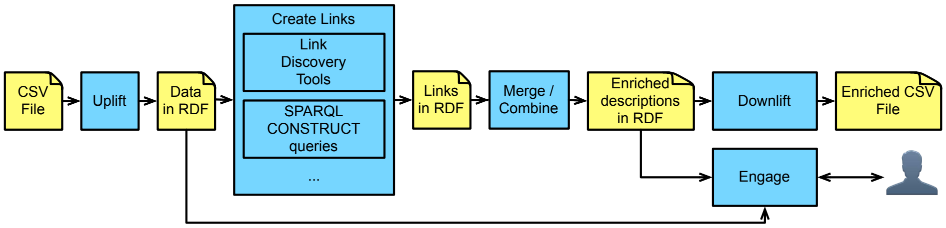


**Figure 1: Method: from uplift and enrichment with a geospatial dimension to downlift and engagement**

datasets are related to each other. For example, Silk has been successfully used to relate Irish place name datasets with geographic datasets [13]. EAGLE [11], on the other hand, adopts Machine Learning techniques (active learning, in this case) to discover how entities are related to each other. Where the former is great if you know how *and* can declare how entities are related (one basically captures that knowledge in a so called "link specification" file), the latter is useful if one wants to discover links based on a (few) exemplars that have been pre-assessed by a user.

With the second approach, SPARQL CONSTRUCT queries allow one to generate RDF graphs based on templates plus the result of the graph pattern in the WHERE clause. An example of a CONSTRUCT query to assert that two counties are disjoint based on a geospatial function in the filter condition is shown in Listing 2. The result of a CONSTRUCT query is an RDF graph that can be stored in any one of the RDF serialization formats available.

How this query creates links between two separate RDF datasets will be explained later. Assuming that the reader understands that merging two RDF files – the CSV data as RDF and the links – is straightforward, we will now proceed to describe the processes of downlift and how to engage with the RDF.

**Listing 2: Example SPARQL CONSTRUCT query.**

```
PREFIX osi: <http://ontologies.geohive.ie/osi#>
PREFIX geo: <http://www.opengis.net/ont/geosparql#">
PREFIX gf: <http://www.opengis.net/def/function/geosparql/>
CONSTRUCT {  ?c1 geo:ehDisjoint ?c2 . }
WHERE {
  ?c1 a osi:County ; geo:hasGeometry ?g1 . ?g1 geo:asWKT ?w1 .
  ?c2 a osi:County ; geo:hasGeometry ?g2 . ?g2 geo:asWKT ?w2 .
  FILTER(gf:sfDisjoint(?w1, ?w2))
}
```

Please note that when creating links, one needs to take care which predicate to adopt in order to express the corresponding relationship between two pieces of data. A plethora of predicates are available in various vocabularies to denote that the entities are the same, similar, related by some spatial coverage, etc. In this paper, we will not advise which might predicate to adopt, as this depends on what a user is attempting to achieve.

## 3.3 Downlift of RDF data into the CSV File

The process of generating non-RDF resources from RDF is called *downlift*. In our approach, we rely on the mapping and the enriched RDF in order to downlift the data into a non-RDF format. The designed algorithm inspects the mapping to recreate the CSV file's structure and appends additional columns for the links that were discovered. When no suitable information for a column name for these new columns is available, a name based on the predicate's URI is used to represent the relationship expressed by the link.

The query in Listing 3 is used in the mapping to retrieve the predicates and the labels from the original CSV file, in the order in which those columns appeared. The labels and their corresponding predicates are kept in lists, which are appended with information about the predicates used to interlink datasets.

**Listing 3: Retrieving the CSV file's structure.**

```
PREFIX rr: <http://www.w3.org/ns/r2rml#>
PREFIX odef:<http://adaptcentre.ie/ont/odef#>
SELECT ?predicate ?label WHERE {
         ?pom odef:label ?label .
         ?pom odef:order ?order .
         ?pom rr:predicate ?predicate .
} ORDER BY ?order
```

Retrieving the, possibly enriched, information about each record is undertaken by generating a query that takes into account the non-existence of values. This is achieved by specifying an optional graph pattern for each predicate. It suffices to state that such a pattern acts as some sort of outer join and does not bind the variables in the pattern when no match is found. Assuming that the previous query returned two predicates (`csv:LAT` and `csv:LONG`) for a particular mapping and `foaf:based_near` was used to interlink datasets, the downlift algorithm will generate the SPARQL query in Listing 4 for retrieving record instances from the RDF. The values bound to variables $?x_0 ... ?x_n$ will then be used to populate the new CSV files.

**Listing 4: Retrieving the new CSV file's data.**

```
PREFIX odef:<http://adaptcentre.ie/ont/odef#>
PREFIX csv: <file:///...# >
SELECT  ?x0 ?x1 ?x2 WHERE {
 ?x a odef:Record .
} ORDER BY ?x
```

## 3.4 Engagement

When it comes to engaging with the data, either for exploring the "raw" data, for creating links with SPARQL CONSTRUCT queries, or for using the enriched RDF, we use Triple Pattern Fragments (TPF) [17], which we have extended with client-side processing of GeoSPARQL [3].[8]

As explained in Section 2.2, TPF consists of a server that can resolve simple queries based on triple patterns, and it is up to the client to compute more complex queries based on those simple result sets. TPF allows a client to consult several servers, which do not need to be hosted at the same domain. We can use this to our advantage by setting up a TPF server for our enriched RDF on one's local machine, and have a client – which can also be run from a user's browser – combine the data with external TPF servers.

---

[8] The implementation of our extension is available at https://github.com/chrdebru/Client.js.

To facilitate this, we generate a TPF server configuration for the user for the output file. If possible, we encourage users to first transform the RDF into a Header-Dictionary-Triples [6] (HDT) file, which is more efficient than loading the local RDF file in memory. When the server is launched, the user can use our extended client.

## 4 DEMONSTRATION

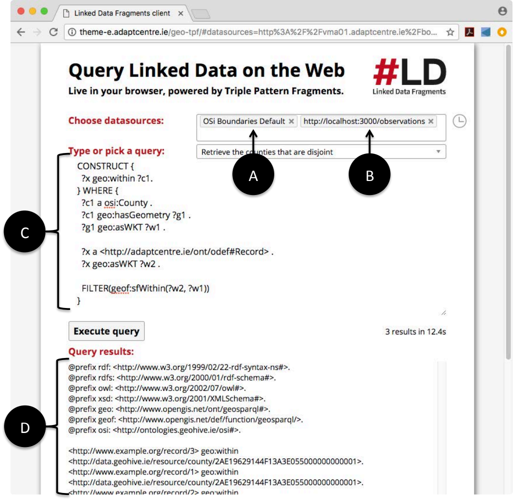
In this section the approach outlined in Fig. 1 is described with reference to a concrete case study we have undertaken in order to demonstrate and evaluate the approach. The study is undertaken in the context of both an ongoing collaboration with the Ordnance Survey Ireland (OSi) – Ireland's national mapping agency – with whom we have published boundary datasets as Linked Data on the Web[9], and the Open Data Engagement Fund, an Initiative by the Irish Government's Department of Public Expenditure and Reform (DPER), that aims to promote the usage of the data made available on Ireland's open data portal[10]. At the time of writing, only a few datasets are currently available on the portal as RDF, with instead the most popular machine-accessible format being Comma-Separated Values (CSV). Our aim in our demonstration was to enrich these CSV files with a geospatial dimension through the use of semantic technologies.

To demonstrate our approach, we will use one of the weather stations datasets available on `data.gov.ie`[11], of which the records are shown in Table 1. The generation of an R2RML mapping with our approach produces the R2RML mapping in Listing 5. Please note that we have omitted a part of the CSV file's namespace and a few predicate-object maps for brevity. The reader will notice that information about the original CSV column are kept using `odef:label` and `odef:order`.

**Listing 5: Generated R2RML mapping.**

```
@prefix rr: <http://www.w3.org/ns/r2rml#> .
@prefix odef: <http://adaptcentre.ie/ont/odef#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix csv: <file:///...# > .
<#TriplesMap>
  rr:logicalTable  [
    rr:sqlQuery  "SELECT rownum() AS ROW_NUM, * FROM
          ↪ fccweatherstationsp201108292221;"
  ] ;
  rr:subjectMap    [
    rr:template "http://www.example.org/record/{ROW_NUM}" ;
    rr:class odef:Record ;
  ] ;
  rr:predicateObjectMap  [
    rr:predicate  csv:LONG ;
    odef:label    "LONG" ;
    odef:order    "5"^^xsd:int ;
    rr:objectMap  [ rr:column  "LONG" ] ;
  ] ;
  # Other predicate-object maps omitted for brevity
  rr:predicateObjectMap  [
    rr:predicate  csv:ROW_NUM ;
    rr:objectMap  [ rr:column  "ROW_NUM" ] ;
  ] .
```

The algorithm and our R2RML engine currently relies on the H2 database engine's CSVREAD[12] function to create a table from a CSV file. We assume that the table name is based on the CSV file's name, column names only consist of letters, digits and underscores, and are case insensitive. Other column names are case sensitive and need to be quoted. This is for column constants (i.e., `rr:column`) not important.



```
CONSTRUCT {
 ?x geo:within ?c1.
} WHERE {
 ?c1 a osi:County .
 ?c1 geo:hasGeometry ?g1 .
 ?g1 geo:asWKT ?w1 .

 ?x a <http://adaptcentre.ie/ont/odef#Record> .
 ?x geo:asWKT ?w2 .

 FILTER(geof:sfWithin(?w2, ?w1))
}
```

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix owl: <http://www.w3.org/2002/07/owl#>.
@prefix xsd: <http://www.w3.org/2001/XMLSchema#>.
@prefix geo: <http://www.opengis.net/ont/geosparql#>.
@prefix geof: <http://www.opengis.net/def/function/geosparql/>.
@prefix osi: <http://ontologies.geohive.ie/osi#>.

<http://www.example.org/record/3> geo:within
<http://data.geohive.ie/resource/county/2AE19629144F13A3E055000000000001>.
<http://www.example.org/record/1> geo:within
<http://data.geohive.ie/resource/county/2AE19629144F13A3E055000000000001>.
<http://www.example.org/record/2> geo:within
```

**Figure 2: Executing a CONSTRUCT query with A) OSi's TPF server, B) the local TPF server, C) the query, and D) the resulting RDF graph.**

In this particular example, we have values for longitude and latitude. We can avail of these properties to create links with OSi's authoritative dataset to discover in which counties, electoral divisions, etc. those weather stations reside by creating a WKT representation of those coordinates. This is done by adding a predicate-object map for generating WKT points, see Listing 6.

**Listing 6: A predicate-object map for WKT points.**

```
rr:predicateObjectMap [
  rr:predicate geo:asWKT ;
  rr:objectMap [
    rr:template "POINT({LONG} {LAT})" ;
    rr:termType rr:Literal ;
    rr:datatype geo:wktLiteral ;
  ] ;
] ;
```

The execution of the R2RML mapping results in a group of triples – Listing 7 contains the triples that are generated for one record in the file – that we can use to setup a local instance of a Triple Pattern Fragments (TPF) server and to use GeoSPARQL in order to discover relationships.

**Listing 7: Part of the generated RDF.**

```
<http://www.example.org/record/2>
  a <http://adaptcentre.ie/ont/odef#Record> ;
  <file:///...#AGENCY> "National Roads Authority" ;
  <file:///...#LAT> "53.4096411069945" ;
  <file:///...#LONG> "-6.22759742761812" ;
  <file:///...#NAME> "M50 Dublin Airport" ;
  <file:///...#ROW_NUM> "2"^^<http://www.w3.org/2001/XMLSchema#int> ;
  <file:///...#WEATHER_READING> "http://www.nratraffic.ie/weather/default.asp?
        ↪ weatherstation=IR57&selected=true&showonly=true" ;
  <http://www.opengis.net/ont/geosparql#asWKT> "POINT(-6.22759742761812
        ↪ 53.4096411069945)"^^<http://www.opengis.net/ont/geosparql#wktLiteral
        ↪ > .
```

Fig. 2 shows how we can execute a query to create links between the records and administrative county boundaries hosted by Ordnance Survey Ireland.[13] In this example, we have chosen to relate the records and boundaries using the `geo:within` relationship

---

[9] http://data.geohive.ie/
[10] http://data.gov.ie/
[11] https://data.gov.ie/dataset/weather-stations
[12] http://www.h2database.com/html/functions.html#csvread

[13] Here we use the LinkedDataFragments jQuery Widget (https://github.com/LinkedDataFragments/jQuery-Widget.js), but with our extended TPF client with support for GeoSPARQL functions.

**Table 1: Records in fccweatherstationsp20110829-2221.csv**

| Name | Weather_Reading | Agency | LAT | LONG |
|---|---|---|---|---|
| M50 Blanchardstown | http://... | National Roads Authority | 53.3704660326011 | -6.38085144711153 |
| M50 Dublin Airport | http://... | National Roads Authority | 53.4096411069945 | -6.22759742761812 |
| Dublin Airport | http://... | Met Éireann | 53.4215060785623 | -6.29784754004026 |

predicate. The RDF graph that is generated from this CONSTRUCT query is stored in a file and can now be used during downlift.

Using the steps outlined in the previous section, we provide the following as input to the downlift algorithm: a) the mapping, b) the RDF that was generated and the RDF containing the links from the CONSTRUCT query (which will be combined), and c) the predicate `geo:within` to create an additional column.

In this example, the column `geo:within` was added containing a reference to the boundary of County Dublin http://data.geohive.ie/resource/county/2AE19629144F13A3E055000000000001. Using this HTTP URI in a browser will redirect a user to an HTML page describing the county. Requesting another format, e.g., "Accept: application/rdf+xml" would have redirected the user to an RDF/XML description of that county. We have thus succeeded in *enriching* the existing CSV file with a spatial dimension.

## 5 DISCUSSION

The importance of geospatial data on the Linked Data Web is not only reflected by the amount of geographic data available in the Linked Data Cloud, but also by the many (commercial) solutions that are available; Oracle Spatial Graph (commercial) solutions and Parliament [1] (academic) are two examples. Our approach does not rely on such bespoke triplestores, but instead we used an extension (see [3]) of Triple Pattern Fragments with client-side processing of GeoSPARQL in JavaScript. We believe that this non-bespoke approach will leverage the uptake of GeoSPARQL for engaging, combining and exploring local and external datasets.

A part of this study was concerned with client-side processing of GeoSPARQL queries. This is handy when no (GeoSPARQL-enabled) SPARQL endpoints are available – a regular occurrence on the Linked Data Web where it is easier to provide Linked Data dumps [17], and avoids a user having to install such solutions. We are, however, aware that computing these queries in JavaScript is not as efficient as relying on bespoke data structures and storage, especially if we run such queries in a browser environment. That said, programmers can avail of the extended TPF Node.js client, which has been integrated in a web application as shown in this study, to build their own applications and hence will provide more control on how to manage resources.

Many of the open datasets that are available on the Irish Government open data portal have no representation in RDF. Though our approach generates R2RML that reflects the structure of the CSV file, it provides a basis for creating meaningful RDF datasets if users extend the mapping by adopting existing vocabularies. This process would look similar to the approach in our demonstrator of adopting GeoSPARQL to represent WKT points. Our goal was to enrich existing non-RDF datasets with a geospatial dimension via RDF, not the creation of meaningful RDF. It is planned, however, to look into this aspect in the future as it will require a more elaborate method to be developed (e.g., based on [5]).

The lack of semantics in CSV files brings us to another point of discussion. Our approach transforms CSV files into RDF to facilitate link discovery; either with tools such as Silk or via SPARQL construct queries. One needs to trust that the one creating the links understands the CSV data and hence whether the discovered links are correct, and whether the links are correctly understood; i.e., the correct RDF predicate to express the relationship was chosen. We deem the latter of particular importance, as the URIs of the chosen predicates will be used to create additional columns, and thus provide a pointer on how to interpret that information in the enriched RDF.

## 6 RELATED WORK

The GeoKnow project, funded by the EU FP7 Programme which ran from 2013 to 2015, aimed to provide a suite of solutions to integrate and enrich external datasets with geospatial data using semantic technologies. The main differences between the GeoKnow project and our study is that the project aimed to develop a suite, called the GeoKnow Generator Workbench [7] which is a tool stack that needs to be installed on a machine and its data lifecycle does not consider the notion of downlift. However they proposed a data lifecycle from extraction to interlinking and analysis. Some of the tools developed in this project are relevant for our study and may be considered for inclusion in the future: TripleGeo [12], LIMES [10], DEER [15], and Facete [16].

TripleGeo allows one to extract features from various standardized formats (such as GML and KML), ESRI shapefiles, and databases with geospatial support. This service can be of use when the CSV data we uplift has information stored in the aforementioned formats. LIMES, like Silk which we referred to in a prior section, is a link discovery framework with support for some geo-spatial similarity metrics. LIMES has also been extended with machine learning techniques for discovering link specifications, as well as approaches to deal with the time complexity when dealing with vast amounts of data [11]. DEER is interesting as it is used to extract implicit geospatial information where they also consider identifying named entities with NLP. Since CSV files may contain descriptions, this could prove to be an interesting approach. DBpedia Spotlight [9] would be another approach to consider for identifying named entities is text and relating those with URIs in DBpedia, which is a Linked Data version of Wikipedia. In order to keep our approach lightweight, this service should be made available to stakeholders in a remote and reliable manner. Finally, Facete is a web-based faceted browsing of RDF geospatial data with functionality to browse data on a map. Facete operates directly on top of a SPARQL endpoint, so we could investigate the integration of Facete with a TPF client.

Our approach is unlikely to compete with bespoke or more elaborate solutions for geospatial data fusion and analysis such as the one proposed in GeoKnow. Instead, our approach aims to be lightweight and is focused on creating and adding a spatial dimension to CSV data, via downlift, and with which stakeholders can engage with Triple Pattern Fragments for exploration and analysis of the data via queries. We have demonstrated that our lightweight approach is sufficient to add location context to CSV files currently available on the Irish Government's open data portal.

## 7 CONCLUSIONS

We presented a lightweight method and set of tools for: transforming non-RDF resources into RDF documents; creating links between RDF datasets; client-side processing of geospatial functions; and transforming (enriched) RDF documents back into a non-RDF format. This study was undertaken in the context of an initiative by the Irish public administration to promote the use of and enhancement of Irish open data. We aimed to add a geospatial dimension to this data using Irish authoritative geospatial datasets provided by the Ordnance Survey Ireland (OSi). The addition of this geospatial dimension would allow one to analyze, link, explore, and even build data analysis applications on top of several datasets using Semantic Web technologies. We illustrated the feasibility of our approach by applying it to one of the CSV datasets available on this portal.

We chose a dataset that was fit for demonstration purposes; it does not reflect the average complexity of datasets on the portal. We furthermore did not provide an example of specifying linkage rules for combining datasets in this paper. Future work would thus include reporting on insights gained on applying the approach to more challenging datasets and linkage rules. Ideally, this would involve setting an experiment for which we can draw different profiles within our network; domain experts in the OSi and the Irish public administration, and computer scientists, for example. Finally, we plan to capture the provenance aspects of the steps proposed in the method as additional metadata to improve transparency, traceability and reproducibility. To achieve this, we will draw inspiration on project-centric provenance information as outlined in [5].

*Acknowledgements.* Within Ordnance Survey Ireland, we would like to thank Éamonn Clinton for his expertise. The ADAPT Centre for Digital Content Technology is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund. This study was also supported by the Irish Government Department of Public Expenditure and Reform's Open Data Engagement Fund initiative.

## REFERENCES

[1] Robert Battle and Dave Kolas. 2012. Enabling the geospatial Semantic Web with Parliament and GeoSPARQL. *Semantic Web* 3, 4 (2012), 355–370. DOI: http://dx.doi.org/10.3233/SW-2012-0065

[2] Christian Bizer, Tom Heath, and Tim Berners-Lee. 2009. Linked Data - The Story So Far. *Int. J. Semantic Web Inf. Syst.* 5, 3 (2009), 1–22. DOI: http://dx.doi.org/10.4018/jswis.2009081901

[3] Christophe Debruyne, Éamonn Clinton, and Declan O'Sullivan. 2017. Client-side Processing of GeoSPARQL Functions with Triple Pattern Fragments. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2017, co-located with 26th International World Wide Web Conference (WWW 2017) (CEUR Workshop Proceedings)*, Sören Auer, Tim Berners-Lee, Christian Bizer, Sarven Capadisli, Tom Heath, Krzysztof Janowicz, and Jens Lehmann (Eds.), Vol. 1809. CEUR-WS.org. http://ceur-ws.org/Vol-1809/article-06.pdf

[4] Christophe Debruyne and Declan O'Sullivan. 2016. R2RML-F: Towards Sharing and Executing Domain Logic in R2RML Mappings. In *Proceedings of the Workshop on Linked Data on the Web, LDOW 2016, co-located with 25th International World Wide Web Conference (WWW 2016) (CEUR Workshop Proceedings)*, Sören Auer, Tim Berners-Lee, Christian Bizer, and Tom Heath (Eds.), Vol. 1593. CEUR-WS.org. http://ceur-ws.org/Vol-1593/article-13.pdf

[5] Christophe Debruyne, Brian Walshe, and Declan O'Sullivan. 2015. Towards a project centric metadata model and lifecycle for ontology mapping governance. In *Proceedings of the 17th International Conference on Information Integration and Web-based Applications & Services, iiWAS 2015, Brussels, Belgium, December 11-13, 2015*, Gabriele Anderst-Kotsis and Maria Indrawan-Santiago (Eds.). ACM, 50:1–50:10. DOI: http://dx.doi.org/10.1145/2837185.2837201

[6] Javier D. Fernández, Miguel A. Martínez-Prieto, Claudio Gutiérrez, Axel Polleres, and Mario Arias. 2013. Binary RDF representation for publication and exchange (HDT). *J. Web Sem.* 19 (2013), 22–41. DOI: http://dx.doi.org/10.1016/j.websem.2013.01.002

[7] Agata Filipowska, Ruben Verborgh, and Axel Polleres (Eds.). 2015. *Joint Proceedings of the Posters and Demos Track of 11th International Conference on Semantic Systems - SEMANTiCS 2015 and 1st Workshop on Data Science: Methods, Technology and Applications (DSci15) 11th International Conference on Semantic Systems - SEMANTiCS 2015, Vienna, Austria, September 15-17, 2015.* CEUR Workshop Proceedings, Vol. 1481. CEUR-WS.org. http://ceur-ws.org/Vol-1481

[8] Robert Isele, Anja Jentzsch, and Christian Bizer. 2010. Silk Server - Adding missing Links while consuming Linked Data. In *Proceedings of the First International Workshop on Consuming Linked Data, Shanghai, China, November 8, 2010 (CEUR Workshop Proceedings)*, Olaf Hartig, Andreas Harth, and Juan Sequeda (Eds.), Vol. 665. CEUR-WS.org. http://ceur-ws.org/Vol-665/IseleEtAl_COLD2010.pdf

[9] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011 (ACM International Conference Proceeding Series)*, Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie N. Lindstaedt, and Tassilo Pellegrini (Eds.). ACM, 1–8. DOI: http://dx.doi.org/10.1145/2063518.2063519

[10] Axel-Cyrille Ngonga Ngomo and Sören Auer. 2011. LIMES - A Time-Efficient Approach for Large-Scale Link Discovery on the Web of Data. In *IJCAI 2011, Proceedings of the 22nd International Joint Conference on Artificial Intelligence, Barcelona, Catalonia, Spain, July 16-22, 2011*, Toby Walsh (Ed.). IJCAI/AAAI, 2312–2317. DOI: http://dx.doi.org/10.5591/978-1-57735-516-8/IJCAI11-385

[11] Axel-Cyrille Ngonga Ngomo and Klaus Lyko. 2012. EAGLE: Efficient Active Learning of Link Specifications Using Genetic Programming. In *The Semantic Web: Research and Applications - 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings (Lecture Notes in Computer Science)*, Elena Simperl, Philipp Cimiano, Axel Polleres, Óscar Corcho, and Valentina Presutti (Eds.), Vol. 7295. Springer, 149–163. DOI: http://dx.doi.org/10.1007/978-3-642-30284-8_17

[12] Kostas Patroumpas, Michalis Alexakis, Giorgos Giannopoulos, and Spiros Athanasiou. 2014. TripleGeo: an ETL Tool for Transforming Geospatial Data into RDF Triples. In *Proceedings of the Workshops of the EDBT/ICDT 2014 Joint Conference (EDBT/ICDT 2014), Athens, Greece, March 28, 2014. (CEUR Workshop Proceedings)*, K. Selçuk Candan, Sihem Amer-Yahia, Nicole Schweikardt, Vassilis Christophides, and Vincent Leroy (Eds.), Vol. 1133. CEUR-WS.org, 275–278. http://ceur-ws.org/Vol-1133/paper-44.pdf

[13] Catherine Ryan, Rebecca Grant, Eoghan Ó Carragáin, Sandra Collins, Stefan Decker, and Nuno Lopes. 2015. Linked data authority records for Irish place names. *Int. J. on Digital Libraries* 15, 2-4 (2015), 73–85. DOI: http://dx.doi.org/10.1007/s00799-014-0129-8

[14] Nigel Shadbolt, Kieron O'Hara, Tim Berners-Lee, Nicholas Gibbins, Hugh Glaser, Wendy Hall, and M. C. Schraefel. 2012. Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems* 27, 3 (2012), 16–24. DOI: http://dx.doi.org/10.1109/MIS.2012.23

[15] Mohamed Ahmed Sherif, Axel-Cyrille Ngonga Ngomo, and Jens Lehmann. 2015. Automating RDF Dataset Transformation and Enrichment. In *The Semantic Web. Latest Advances and New Domains - 12th European Semantic Web Conference, ESWC 2015, Portoroz, Slovenia, May 31 - June 4, 2015. Proceedings (Lecture Notes in Computer Science)*, Fabien Gandon, Marta Sabou, Harald Sack, Claudia d'Amato, Philippe Cudré-Mauroux, and Antoine Zimmermann (Eds.), Vol. 9088. Springer, 371–387. DOI: http://dx.doi.org/10.1007/978-3-319-18818-8_23

[16] Claus Stadler, Michael Martin, and Sören Auer. 2014. Exploring the web of spatial data with facete. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014, Companion Volume*, Chin-Wan Chung, Andrei Z. Broder, Kyuseok Shim, and Torsten Suel (Eds.). ACM, 175–178. DOI: http://dx.doi.org/10.1145/2567948.2577022

[17] Ruben Verborgh, Miel Vander Sande, Olaf Hartig, Joachim Van Herwegen, Laurens De Vocht, Ben De Meester, Gerald Haesendonck, and Pieter Colpaert. 2016. Triple Pattern Fragments: A low-cost knowledge graph interface for the Web. *J. Web Sem.* 37-38 (2016), 184–206. DOI: http://dx.doi.org/10.1016/j.websem.2016.03.003