Using the Linked Logainm Dataset

Rebecca Grant
Nuno Lopes
Eoghan Ó Carragáin
Catherine Ryan


dri
Digital Repository of Ireland
Taisclann Dhigiteach na hÉireann

DERI

DCU

nli
Leabharlann
Náisiúnta
na hÉireann
National Library
of Ireland

An Roinn
Ealaíon, Oidhreachta agus Gaeltachta
Department of
Arts, Heritage and the Gaeltacht

Linked Logainm is a collaborative project led by the Digital Repository of Ireland (DRI), the Digital Enterprise Research Institute (DERI), Fiontar at Dublin City University (DCU), and the National Library of Ireland (NLI) and the Placenames Branch of the Department of Arts, Heritage and the Gaeltacht.

The Linked Logainm project has created a Linked Data version of the authoritative bilingual database of Irish place names, logainm.ie, which was developed by Fiontar in collaboration with the Placenames Branch of the Department of Arts, Heritage and the Gaeltacht.

The project report gives an overview of Linked Data technologies and the processes undertaken to transform the logainm.ie dataset to Linked Data and to connect it to the LinkedGeoData, GeoNames and DBpedia data sets.

The Location LODer[1] demonstrator website gives an interactive introduction to the potential of the Linked Logainm concept, using logainm.ie data and a map interface to allow users to explore content from sources, including Europeana, the NLI and Wikipedia.

This document suggests other potential uses for the Linked Logainm data set.

---

[1] http://apps.dri.ie/locationLODer (22 October 2013).

When citing or attributing this work, please use the following: Grant, R., Lopes, N., Ó Carragáin, E., Ryan, C. (2013),'Using the Linked Logainm dataset'. Dublin: Royal Irish Academy and National Library of Ireland; Galway: NUI Galway.

DOI: 10.3318/DRI.LODer.2013.3
Digital Repository of Ireland series, no. 4

Who is this document for?

The Linked Logainm dataset is of potential use to any person, project or institution aiming to make content relating to Irish places available on the Web. The publication of logainm.ie data in a structured, computer-readable format allows its value to be reused by computer scientists, web developers, the heritage community and information professionals.

This document provides use cases and examples for those who are interested in working with the Linked Logainm dataset, and who have some technical experience. We have split the examples below into those that are more relevant to data curators and information professionals, and those of interest to computer scientists and developers.

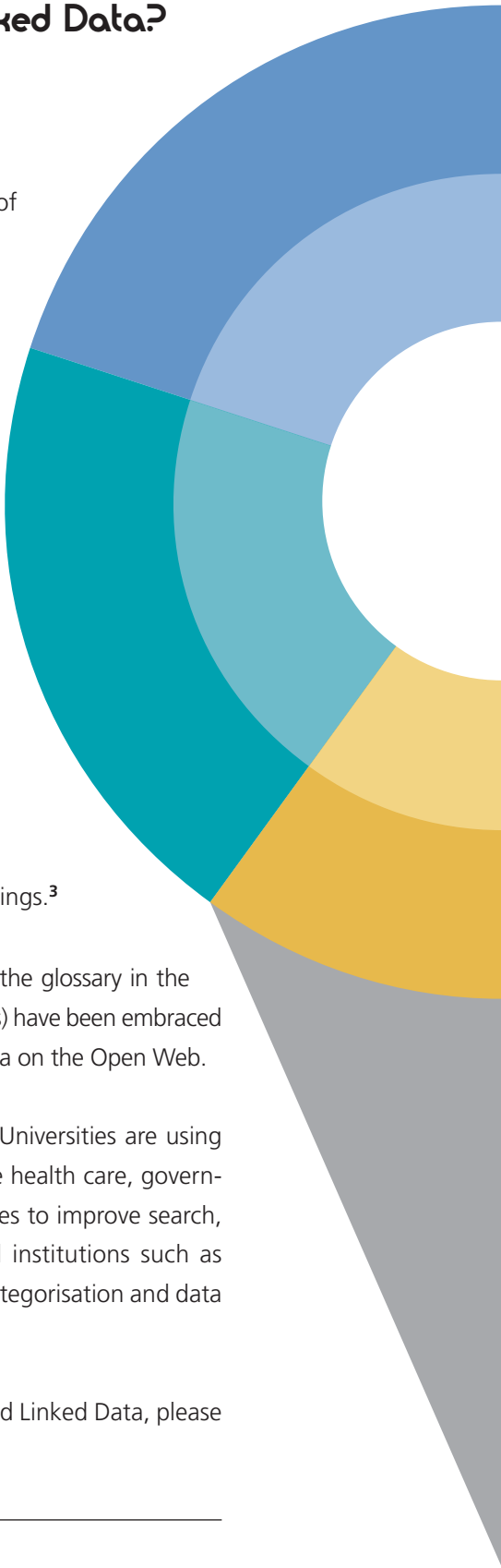
Data curators and information professionals

In the examples given below, we list software tools such as OpenRefine and Pundit that provide excellent end-user interfaces and allow data curators and information professionals to work directly with data from Linked Logainm (referred to as 'End-user tools' below).

Computer scientists, developers and data publishers

In order to access the data directly, Linked Logainm can be used by computer scientists and developers familiar with Linked Data technologies such as SPARQL and RDF (referred to as 'Developer tools' below).

What is Linked Data?



'Linked Data is to spreadsheets and databases what the Web of hypertext documents is to word processor files.'[2]

It refers to data published on the Web following a set of principles designed to lower the barriers to linking data silos by building on the core, proven technologies of the Web. These principles are:

- Use URIs (Unique Resource Identifiers) as names for things (e.g. places, people, concepts);
- Use HTTP URIs so that people can look up those names (just like they look up a web page using a HTTP address);
- Provide useful information when someone looks up a URI using the Linked Data standards (RDF, SPARQL);
- Include links to other URIs so that people can discover more things.[3]

These principles and technologies (URIs, RDF, HTTP and RDFS, see the glossary in the appendix at the end of this document for explanations of these terms) have been embraced internationally as best practice means of publishing and relating data on the Open Web.

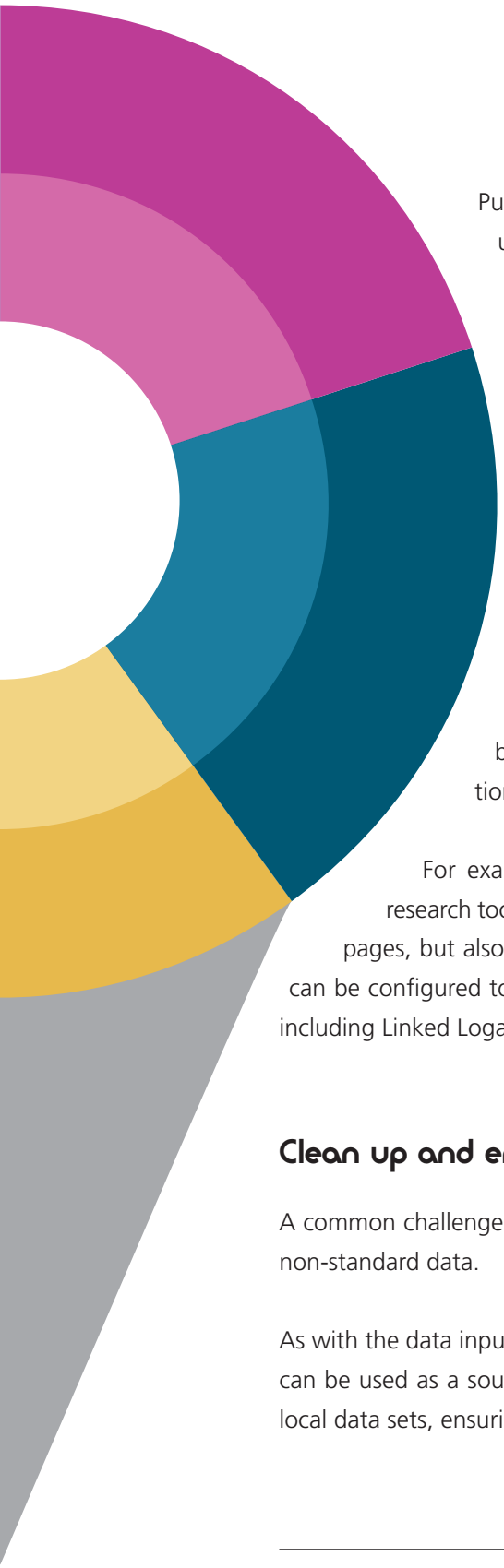
Numerous sectors have implemented Linked Data technologies. Universities are using Linked Data to improve education and learning technology, while health care, government, energy, IT and eTourism sectors are using these technologies to improve search, data integration, content management and discovery. Cultural institutions such as Europeana and the BBC rely on Linked Data to improve content categorisation and data integration, and to enhance search and content discovery.[4]

For more information about the principles and technologies behind Linked Data, please see the appendix to this document.

---

[2] http://www.w3.org/wiki/LinkedData (22 October 2013).
[3] http://www.w3.org/DesignIssues/LinkedData (22 October 2013).
[4] An overview of different Linked Data data sets and their breakdown by different categories and domains can be found at http://lod-cloud.net/state/.

End-user tools



Publishing data on the Open Web allows people to reuse it in unforeseen ways. A few possibilities for such data use drawn from Linked Data utilisations elsewhere are given below.

Controlled vocabulary for data input

The logainm.ie data set provides a unique source for standard, authorised forms of Irish place names in both English and Irish. As this data set is now available in a structured form on the Open Web, it could be used as a source for creating Irish place name entities in local data sets or metadata records.

A potential application could involve developing a look-up service based on the Linked Logainm data set, which would allow integration into data input forms for local applications.

For example, Pundit[5] is an award-winning, open-source collaborative research tool, which allows users 'not only to comment, bookmark or tag web pages, but also to create semantically structured data while annotating'. Pundit can be configured to retrieve Linked Data entities from different SPARQL end points, including Linked Logainm.

Clean up and enhancement of legacy data or metadata

A common challenge when integrating data sets is the need to correct inconsistent or non-standard data.

As with the data input example above, the Linked Data version of the logainm data set can be used as a source against which to automatically compare Irish place names in local data sets, ensuring they match the standard form in English or Irish.

[5] http://www.thepund.it (22 October 2013).

Developers can write code to directly access the data, or make use of free tools like OpenRefine,[6] which are able to 'reconcile' local data sources with Linked Data resources like Linked Logainm.

To add the Logainm RDF reconciliation service in OpenRefine, users need to navigate to 'RDF' > 'Add reconciliation service' > 'Based on SPARQL endpoint...', and fill in the following information:

Name: (any name for the service)
Endpoint URL: http://data.logainm.ie/sparql
Type: Virtuoso
Label Properties: Also check 'foaf:name'

Enhance search and discovery systems

By relating data to Linked Logainm and by keeping a reference to the Linked Logainm URI, a local system could access the unique information that is being developed by Fiontar and the Placenames Branch of the Department of Arts, Heritage and the Gaeltacht, including:

- Information on the equivalence of Irish and English forms of Irish place names;
- Information on alternate, superseded or non-approved spellings and versions of Irish place names;
- Information on the hierarchical relationship between counties, baronies, townlands, parishes and other features of Ireland.

This additional information could act as a powerful knowledge base for search and discovery applications, where users are provided with search suggestions or redirections based on structured relationships found in the Linked Logainm data set.

[6] http://openrefine.org (22 October 2013).

Publishing information about Irish place names: a linking hub for URIs

In addition to publishing data using URIs, HTTP and RDF, the Linked Data principles encourage publishers to relate their data to URIs from other open data sets.

Accurately linking data sets can be a time-consuming and costly exercise even with automated tools, which means that publishers must be selective. In practice, larger open data sets such as DBpedia[7], Freebase[8] and GeoNames[9] have emerged as linking hubs for the Linked Data cloud. By creating a single link to these common hubs, publishers can relate their data to hundreds or thousands of other data sets.

This was the approach taken in the Linked Logainm project, where logainm.ie data was connected to DBpedia and GeoNames, thereby linking it to all other data sets which reference the same URIs from those resources.

One of the motivating factors for the Linked Logainm project was that not all Irish place names, especially at the barony, parish and townland level are represented in data sets such as DBpedia and GeoNames. Linked Logainm is therefore a very useful target when publishing data sets which reference Irish place names at a granular level, since there is a greater chance that these entities will be present. Of course, since Linked Logainm already contains links to DBpedia and GeoNames, by matching resources to Linked Logainm, publishers automatically get the benefit of those links too.

In this way, Linked Logainm has the potential to act as localised hub for Irish place name URIs and a gateway to the wider Linked Data cloud.

Silk[10] and LIMES[11] are two of the available tools that help with creating links between different RDF datasets. In creating the Linked Logainm data set we used the Silk framework to determine the links between Linked Logainm and DBpedia, LinkedGeoData and GeoNames. Triplestores may include the ability to publish Linked Data based on its data, for example Openlink Virtuoso.[12] Alternatively, other tools like Pubby,[13] allow you to publish Linked Data based on any available SPARQL end point.

---

[7] http://wiki.dbpedia.org/About (22 October 2013).
[8] http://www.freebase.com/ (22 October 2013).
[9] http://www.geonames.org/ (22 October 2013).
[10] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/ (22 October 2013).
[11] http://aksw.org/Projects/LIMES.html (22 October 2013).
[12] http://virtuoso.openlinksw.com/ (22 October 2013).
[13] http://wifo5-03.informatik.uni-mannheim.de/pubby/ (22 October 2013).

Querying information about Irish place names

Since 2008, researchers have been able to query the uniquely rich source of information about Irish place names through the logainm.ie website. By publishing the raw data behind logainm.ie as Linked Data, developers can now access standard Linked Data technologies and tools to interrogate the relationships that exist in the Logainm data set and to enhance their own data and applications.

The structured Linked Logainm data from the *logainm.ie* database can be accessed and interrogated at http://data.logainm.ie/sparql/, using the SPARQL query language. Structured information about each place name can also be retrieved in different serialisations (e.g. XML, N3, and JSON) by accessing the respective identifier using content negotiation.

The Linked Logainm data is described using the NeoGeo vocabulary,[14] where every place name is defined as being of type 'Feature' and the respective geographical coordinates are a distinct entity related by the 'Geometry' property.

For instance, the Linked Logainm URI identifying the City of Dublin is http://data.logainm.ie/place/1375542 while the URI for its geographical coordinates is http://data.logainm.ie/geometry/1375542. Each place name is also assigned a type specific to the logainm.ie data set (for example barony, town, townland, etc.). The names are associated with each place via the 'foaf:name' property, and annotated with a respective language tag ('en' or 'ga').

Accessing Linked Data with SPARQL

Information resources described using RDF, and expressed in RDFS, OWL, SKOS and others, are saved to database management systems known as 'triplestores'. In the same way that relational databases use SQL to query tables in the database, triplestores use SPARQL (SPARQL Protocol and RDF Query Language)[15] to query and retrieve information stored in RDF triple format.

SPARQL end points are entry points for querying triplestores using SPARQL queries, which are structured as follows (using examples from Linked Logainm's SPARQL end point (http://data.logainm.ie/sparql)):

---

[14] http://geovocab.org/doc/neogeo/ (22 October 2013).
[15] http://www.w3.org/TR/rdf-sparql-query/ (22 October 2013).

Basic query structure

```
select * where { ?s ?p ?o }
```

In the above example, 'select' is the type of query we are submitting. The asterisk(*) indicates that we want all variables returned. The items in the curly brackets stand for the subject, predicate and object.

Select all subjects, predicates and objects where the name of the subject is Dublin [results]]

```
select * where { ?s ?p ?o . ?s foaf:name "Dublin"@en}
```

In the above example, we are querying all subjects, predicates and objects where the name of the subject is Dublin (in English).

Select all place names of type "county" [results]

```
select ?s where { ?s rdf:type <http://data.logainm.ie/category/CON> }
```

Select the Irish and English names of "Dublin" and its geographical coordinates [results]

```
select ?nameGA ?nameEN ?lat ?long
where { <http://data.logainm.ie/place/1375542> foaf:name ?nameEN, ?nameGA;
                                        <http://geovocab.org/geometry#geometry> ?g
.
        ?g geo:lat ?lat ; geo:long ?long .
        filter(lang(?nameEN) = "en" && lang(?nameGA) = "ga").
}
```

Select the URIs of "Dublin" in other data sets [results]

```
select ?uri where { <http://data.logainm.ie/place/1375542> owl:sameAs ?uri . }
```

Some tutorials on the SPARQL language can be found at:

- http://learningsparql.com/ (22 October 2013);
- http://jena.apache.org/tutorials/sparql.html (22 October 2013);
- http://www.linkeddatatools.com/querying-semantic-data (22 October 2013);
- http://www.cambridgesemantics.com/semantic-university/sparql-by-example (22 October 2013).

Accessing Linked Data via dereferencing URIs

Structured information about each place name can also be retrieved, in different formats, from the respective identifier via content negotiation. A common tool for accessing data in a language independent manner is cURL.[16] For example, the URI identifying the city of Dublin is http://data.logainm.ie/place/1375542, and we can use cURL to access the different representations:

```
curl -H "Accept: text/rdf+n3" http://data.logainm.ie/place/1375542
```

Different types of HTTP accept headers can be:

- 'text/plain' for the NTriples representation;
- 'text/rdf+n3' for the Turtle representation;
- 'application/rdf+json' for the JSON representation;
- 'application/rdf+xml' for the XML representation.

This information can also be accessed pragmatically from different languages, for instance, in Javascript the JSON results can be retrieved with JQuery:

```
$.getJSON('http://data.logainm.ie/place/1375542', callbackFunction )
```

From Python similar results can be achieved using the 'requests' library (http://docs.python-requests.org/en/latest/):

```
import requests
requests.get("http://data.logainm.ie/place/1375542", headers = { 'Accept':
'application/rdf+json})
```

[16] http://curl.haxx.se/ (22 October 2013).

Appendix:
Linked data glossary

URIs

Uniform Resource Identifiers (URIs) are used to provide unique names for resources on the Web. They can be used to identify and express subjects, relationships, properties, values or anything else existing on the Web, as well as 'real world objects' such as people and places. Using a URI instead of a string of text to identify an entity or resource removes any ambiguity between people or places that have the same name, errors in naming entities due to a misspelling or misplaced or absent punctuation, for example many library management systems currently consider 'James Joyce' and 'James Joyce.' to be different entities. Instead we can use a URI for James Joyce[17] as identifier.

RDF

The Resource Description Framework (RDF)[18] is a graph data model developed by W3C for representing and exchanging information on the Web. RDF makes statements, called 'triples', and they take the form subject, predicate and object, where the subject is the entity or resource and the object is another resource or value. The predicate is the relationship between the two entities and is defined using predefined vocabularies. By combining any number of these statements, we create a network of triples (RDF graph). RDF requires that URIs are used to name things and relationships and, by doing so, this data can be understood by computers, is persistent and unambiguous, and can be shared across the Web. RDF has also been serialised in XML and other formats.

---

[17] For instance the DBpedia URI http://dbpedia.org/resource/James_Joyce (22 October 2013).
[18] http://www.w3.org/RDF/ (22 October 2013).

RDF schema

While RDF is the basic data model underlying Linked Data and the Semantic Web, it does not define the relationships in these statements and the terms used to express them: this task is left to other languages. RDFS is the most basic knowledge representation language providing a basic core type system of classes and properties and indicates how they should be used together.[19]

Web Ontology Language (OWL)

The Web Ontology Language (OWL) is a more extensive and expressive knowledge representation language than RDFS and is used primarily as a language for creating and expressing ontologies on the Web. It is used to describe and define terms within a particular domain of interest or subject and the relationships between them.

Many ontologies have been developed using a combination of RDFS and OWL as their base languages. Examples of this include the Friend of a Friend (FOAF), a social ontology for describing people, their activities and the relationships between them and the Simple Knowledge Organisation System (SKOS), another standard for the organisation of knowledge on the Web.

[19] http://www.w3.org/TR/rdf-schema/ (22 October 2013).