Integrating Public Datasets Using Linked Data: Challenges and Design Principles

Tope Omitola[1], Christos L. Koumenides[1], Igor O. Popov[1], Yang Yang[1], Manuel Salvadores[1], Gianluca Correndo[1], Wendy Hall[1], and Nigel Shadbolt[1]

Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
{ t.omitola, clk1v07, ip2g09, yy1402,
ms8, gc3, wh, nrs } @ecs.soton.ac.uk

Abstract. The world is moving from a state where there is paucity of data to one of surfeit. These data, and datasets, are normally in different datastores and of different formats. Connecting these datasets together will increase their value and help discover interesting relationships amongst them. This paper describes our experience of using Linked Data to inter-operate these different datasets, the challenges we faced, and the solutions we devised. The paper concludes with apposite design principles for using linked data to inter-operate disparate datasets.

1 Introduction

The dramatic changes in production, distribution, aggregation, interpretation, and consumption of data have triggered the consumer data revolution. Traditionally, paid specialists actively collected data for a specific purpose. Now, individuals stream their attention, intention, location, etc., and explicitly share highly sensitive data, including their DNA and financial information. A second parallel revolution is simultaneously occurring, and it is the continued growth in the adoption and usage of ontologies and semi-structured data in government and industries. Increasingly, private corporations and governments are realising the potential of encoding knowledge as (semi) structured data. Facebook[1] recently announced their Open Graph Protocol[2] which allows a Facebook user to integrate other non-Facebook web pages into the users social graph. Open Graph Protocol uses machine-processable semi-structured data to mark up web pages. Various governments, in order to improve the delivery of services to their citizens, are opening up their data and publishing these data in semi-structured format, many of them in RDF, to improve the delivery of goods and services. The United States government has set up *data.gov*[3] to release public data. The UK

[1] www.facebook.com

[2] http://developers.facebook.com/docs/opengraph

[3] data.gov

Government, keen to unlock the benefits of economic and social gain of public sector information (PSI) reuse, has set up *data.gov.uk*.

With these datasets in different knowledge bases and data stores, the biggest challenges and opportunities lie in connecting these disparate datasets to create new sets for analysis, and to discover interesting patterns and relationships. Data integration has been a problem since databases have existed, but the amount of potentially relevant data available to a researcher or a curious individual is now thousands of times larger - the problem has moved from enterprise to the mainstream.

In this paper, we describe an implementation of using open geographical data as a core set of "join point"(s) to integrate, and inter-operate between, different public datasets. We describe the challenges faced during the implementation, which include, sourcing the datasets, publishing them as linked data, and normalising these linked data in terms of finding the appropriate join points from the individual datasets, as well as developing the client application used for data consumption. We describe the design decisions and our solutions to these challenges. We conclude by drawing some general principles from this work.

2 Application Case Study

2.1 Introduction

We investigated the use of disparate sets of data in an effort to better understand the challenges of their integration using Semantic Web approaches. Part of this investigation involved ascertaining the datasets that were available, their formats, and converting them into (re)usable formats, asking our questions, and also linking our data back into the linked data cloud[4]. The issue we started with was how to deal with linked data that are centred around the democratic system of political representation in the United Kingdom. We noticed that a lot of UK governmental data are already referenced by geography. The Ordnance Survey has produced a number of ontologies and an RDF data set that represents the key administrative entities in the UK [3]. The questions we asked were what kinds of problems will be encountered from developing a service that uses an administrative entity, i.e. geography data, linked with other data, such as criminal statistics data, the Members of Parliament of these entities (their data), the mortality rates of these entities, and the National Health Service (NHS) hospital waiting times.

2.2 Design Decisions

1. Sourcing the datasets. Since many of the datasets of interest were not yet in linked data format, we could not take advantage of the automatic resource discovery process as enunciated in [10]. We sourced the data by going to the relevant department of government websites. Some datasets were in PDF

[4] http://linkeddata.org

and HTML formats, while some were in XLS formats. For reasons of data fidelity, ability to source from a wider range of public sector domains, and to have increased value that comes from many information linkages, we chose the ones in XLS formats. In future, we do expect many of these datasets to be sourced via the U.K. government's public datastore[5]. This should aid the discovery process of consuming (linked) data.

2. Selection of RDF as the normal form: We decided to use RDF as the normal form for the datasets. RDF offers many advantages, such as provision of an extensible schema, self-describing data, de-referenceable URIs, and, as RDF links are typed, safe merging (linking) of different datasets. We chose the RDF/Turtle representation of RDF triples for its compactness and clarity.
3. We chose a central 4store [4] system to store and manage our RDF triples. 4store provides a robust, scalable, and secure platform to manage RDF triples[6].
4. Modelling multidimensional data. The real world is complex, multidimensional (of space and time) and multivariate, and so are our chosen datasets. They contain dimensions such as time, geographical regions, employment organisations, etc. To model this multi-dimensionality, we chose SCOVO[5]. SCOVO is an expressive modelling framework for representing complex statistics.
5. Many of the datasets we used have notions of geography or region. To join them together, we used geographical location data as the set of "join point"(s). We chose to use the Ordnance Survey's geographical datasets [3] as our set of join points. The Ordnance Survey datasets are relatively stable and fairly authoritative.
6. Although our datasets had concepts of geography, the names given to particular geographical regions differ. As many of these regions refer to the same geographical boundaries, we used `owl:sameAs` to assert equivalences between them.
7. Consumption of data. We used Exhibit[7] to develop the client application. Exhibit allows quick development of Web sites that support various data-centric interactions such as faceted browsing and various representation formats over data such as tables, timelines, thumbnail views, etc.

3 Public Sector DataSets - Publication and Consumption

3.1 Datasets

We used five major datasets. Table 1 lists the data sets used, their formats, and a brief description of the data. They include datasets of Members of Parliament (MPs), Lords, their corresponding constituencies and counties, relevant websites,

[5] http://data.hmg.gov.uk/about

[6] As of 2009-10-21 it's running with 15B triples in a production cluster to power the DataPatrol application(*http://esw.w3.org/topic/LargeTripleStores*)

[7] http://simile.mit.edu/wiki/Exhibit/API

MPs' expenses and votes, and statistical records about crime, hospital waiting time, and mortality rates.

Data Source	**Format**	**Dataset**
Publicwhip.org.uk	HTML	MP Votes Records, Divisions, Policies
Theyworkforyou.com	XML Dump	Parliament, Parliament expenses
Homeoffice.gov.uk	Excel Spreadsheet	Recorded crime (English and Wales 2008/09)
Statistics.gov.uk	Excel Spreadsheet	Hospital Waiting List Statistics (English 2008/09)
Performance.doh.gov.uk	Excel Spreadsheet	Standardised mortality ratios by sex (English and Wales 2008)
Ordancesurvey.co.uk	Linked Data	National mapping agency, providing the most accurate and up-to-date geographic data

Table 1. Targeted Government data sources, formats, and description of dataset.

3.2 Modelling the datasets

Most of our vocabularies come from Friend-of-a-Friend (FOAF)[8], Dublin Core (DC)[9], and SCOVO, thereby following the advice given in [9] to re-use terms from well-known vocabularies.

Modelling the Hospital Waiting List. Each row, of this dataset[10], consisted of data for each health care provider, or a National Health Service (NHS) Hospital Trust, in England and Wales. The columns consisted of various data that were of no interest to us. One of the columns, "Patients waiting for admission by weeks waiting", was made up of several columns which had data for patients waiting for hospital operations, and each of these columns was divided into weekly waiting times, from those waiting between 0 and 1 week, continuing to those waiting for more than thirty weeks. We modelled the data as follows. The time period, 2008/09, the NHS Hospital Trust (e.g. South Tyneside NHS Foundation Trust), and the waiting periods are `scovo:Dimension`(s). The value for patients that had been waiting for hospital operation from between zero to one week at South Tyneside NHS Foundation Trust is modelled as:

```
:A_RE9 rdf:type waitt:OrgName; dc:title "RE9";
    rdfs:label "South Tyneside NHS Foundation Trust";
    statistics:SHA "Q30"; statistics:org_code "RE9".
```

[8] http://xmlns.com/foaf/spec/

[9] http://dublincore.org/documents/dcmes-xml/

[10] http://www.performance.doh.gov.uk/waitingtimes/index.htm

```
:ds1_1_2 rdf:type scovo:Item; rdf:value 185;
    scovo:dataset :ds1; scovo:dimension :w0to01week;
    scovo:dimension :A_RE9; scovo:dimension :TP2008_09.
```

Modelling the UK Parliament. The datasets of information for MPs, Lords, constituencies, counties, MPs' expenses and votes were downloaded from the Parliament Parser[11]. Most of these were raw XML files. The Parliament Parser provides structured versions of publicly available data from the UK parliament. Members of parliament and lords were modelled as `foaf:person`(s) with parliament identities corresponding to their roles in the Parliament at different time periods. Constituencies and counties were embodied as `dc:jurisdiction`(s) and linked to their corresponding MP and Lord identities via the `dc:coverage` property. Political parties and the Houses themselves were in turn modelled as `foaf:group`(s), while MPs expenses and votes were modelled using the SCOVO ontology.

3.3 Converting Datasets to RDF

Most of our data were in spreadsheet or comma-separated-values (csv) formats. There are inherent problems with re-using data published in spreadsheet format. These include:

1. little or no explicit semantic description, or schema, of the data. An example of this can be seen from the Hospital Waiting List where there were codes given names such as "SHA Code", and "Org Code", without explanation of their relationships with the rest of the data in the spreadsheet.
2. more difficult to integrate, or link, data from disparate data sources. An example of this can be seen from the Home Office data where each area's value for a crime was given. It will be good to know how this data was arrived, and linking it with the data sources from whence they come would have been useful (e.g. for provenance and validation).

We developed a number of scripts to automatically convert the spreadsheets' data, and used the Jena Semantic Web Framework[12] to convert the Parliament data, into RDF triples. These triples were stored in our local 4store system.

3.4 Alignment of the Datasets

The process of aligning the datasets relied on the correct identification of `owl:sameAs` relations between the geographic concepts of the datasets and the corresponding relevant entities in the Ordnance Survey Administrative Geography (OS Admin Geo). The relevant entities here were constituencies data from the Ordnance Survey (OS).

[11] http://ukparse.kforge.net/parlparse/

[12] http://jena.sourceforge.net/

Aligning the Hospital Waiting List The geographical entity here was the full name of each NHS Trust in England and Wales, e.g. "South Tyneside NHS Foundation Trust". We employed the Google Maps API[13] to get the locations of these NHS Trusts. The Google Maps API returned for each geographical entity, with increasing precision, the Administrative Area, Sub Administrative Area, Locality, as well as their lat/long coordinates. We then manually queried the OS, using string matching, for the constituency names of this entity. In case the string matching operation failed, we queried TheyWorkForYou API, for the constituencies, giving it the lat/long values.

Aligning the UK Parliament Data In the United Kingdom, boundaries of constituencies and constituency names change every few years. This affected the precise alignment of the data from the Parliament Parser and the OS Admin Geo. The Parliament Parser solves the problem of constituencies changes by assigning a new identifier to it[14]. The Ordnance Survey, however, only defines constituencies according to their latest classification by the UK Parliament. Therefore, only a partial alignment of these two datasets was possible.

3.5 Linked Data Consumption

The application scenario we envisaged is as follows: a user wants to find out some information about their geographical region - political, social etc. They have no knowledge, however, of the kind of data they might find nor are they knowledgeable in all the various geographical entities their place of residence is part of. The application acts as an aggregator of information based on the user's postal code, which they input at the start of the application. The application then tries to generate views of data from different topics along with widgets that allow the user to further explore the data retrieved.

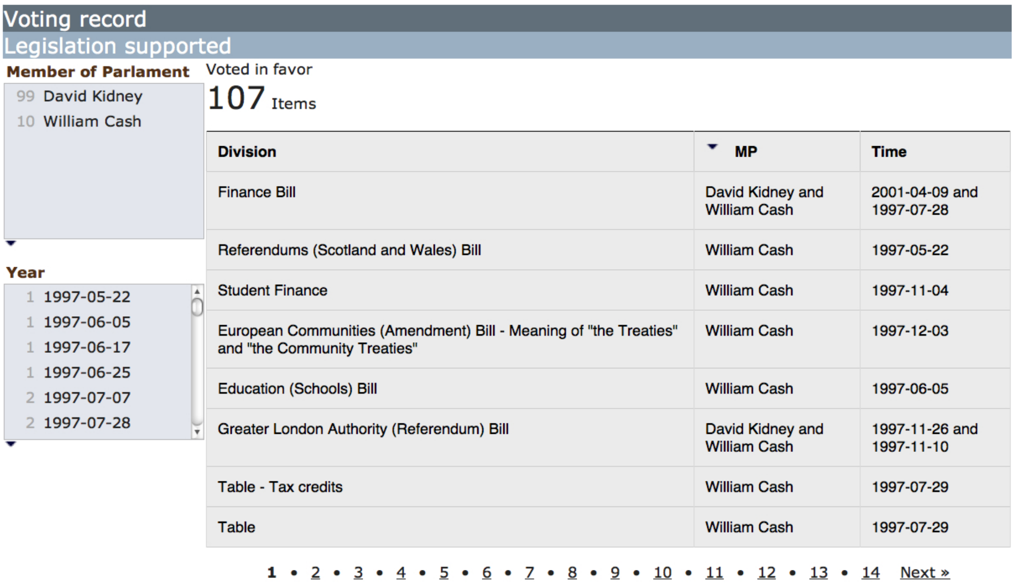
In the application[15], the geographical region acts as the context for the displayed data, and is the central point from which the application follows links to find the data to display. For example, the application starts off showing political information, such as the political representatives for that area. Since in our data the constituency is the lowest geographic entity in the hierarchy, it starts by showing the political representation for that constituency. The interface shows the different MPs that have served or are in office for that constituency, plots a timeline view of their terms in office, and shows data about their voting records (Figure 1), and expenses. Additionally, the application generates facets for the user to quickly filter through the information. To keep the load of information low and present relevant information, we restrict the application to presenting data for MPs for the last two decades. We note that the application accounts

[13] http://code.google.com/apis/maps/

[14] "Unique identifiers and alternative names for UK parliamentary constituencies. A constituency is given a new id whenever its boundaries change." [see http://ukparse.kforge.net/svn/parlparse/members/constituencies.xml]

[15] The application is at *http://psiusecase.enakting.org/*

for any temporal inconsistencies by matching the time periods between these aggregated data, as there are cases where a certain MP had served in two different constituencies and data on their expenses are available only for one of those terms. In such cases information is restricted only to that which applied for the period they are in office for the constituency currently under view. Going up the hierarchy, at each level the application tries to find data about different topics. For example, it finds the county which contains the constituency and tries to retrieve crime data for that county. If it finds data, it displays them.



Voting record

Legislation supported

Member of Parlament

99 David Kidney
10 William Cash

Year

1 1997-05-22
1 1997-06-05
1 1997-06-17
1 1997-06-25
2 1997-07-07
2 1997-07-28

Voted in favor

107 Items

Division	MP	Time
Finance Bill	David Kidney and William Cash	2001-04-09 and 1997-07-28
Referendums (Scotland and Wales) Bill	William Cash	1997-05-22
Student Finance	William Cash	1997-11-04
European Communities (Amendment) Bill - Meaning of "the Treaties" and "the Community Treaties"	William Cash	1997-12-03
Education (Schools) Bill	William Cash	1997-06-05
Greater London Authority (Referendum) Bill	David Kidney and William Cash	1997-11-26 and 1997-11-10
Table - Tax credits	William Cash	1997-07-29
Table	William Cash	1997-07-29

1 • 2 • 3 • 4 • 5 • 6 • 7 • 8 • 9 • 10 • 11 • 12 • 13 • 14 Next »

Fig. 1. Displaying voting record data for selected British Members of Parliament.

4 Conclusions and Future Work

In an effort to provide greater transparency amongst public sector departments and to target public services to areas of best need, governments are actively opening up public data. However much of these data are in non-linked formats. The data models are difficult to understand and re-use, and closed to web-scale integration. Publishing these data in linked data format would make it easier for them to be re-useable and interlinked.

In this work, we took data from disparate public data sources converting them into linked data format using geographical data, as the set of “join point”(s), to compose them together to form a linked integrated view. Several issues and challenges needed to be solved to build an integrated view of these disparate datasets.

4.1 Challenges in Data Publication

1. Although there is an increase in the amount of public data being made available, there is still a paucity of data in the right formats. Most data are still in HTML, PDF, and XLS formats, publishing and re-publishing these data in linked form will be very useful,
2. Many of these disparate datasets may not cover the same temporal intervals. This may make comparison over time complex. Most of the missing data are likely to be stored in hard-to-reach places in their respective government departments. As more of them are published, future temporal interval misalignments will be mitigated,
3. Data/Instance (Ontology) Alignment. Whenever there is more than one way to structure a body of data, there will be data and semantic heterogeneity when they are joined together. Because they are more flexible, semi-structured data exacerbates this problem, and as there will be more of them, the linked data cloud will add to this. Various mechanisms have addressed these problems. Semi-automatic mechanisms use an admixture of human and software, e.g. see [12], while fully automated methods, such as [13], aim to discover data and schema overlaps with no human intervention. We did not use any automatic methods in this exercise, and used mainly manual methods. Linking datasets required us to resort to string matching. This method is certainly unscalable to hundreds and thousands of linked datasets that are expected to come on stream in the next few years.

4.2 Challenges in Data Consumption

1. One of the biggest challenges we experienced was the low level of interoperability between the user interface (UI) and the underlying data. This meant that data consumption was not direct and needed to be converted and re-modelled in order to be shown in the UI. This conversion, however, was not straightforward as we had to frequently query the store and use a proxy to construct an entirely new data model for the UI. This highlights two important issues pertaining to UI over heterogeneous data:
   (a) The lack of UIs to quickly browse, search or visualise views on a wide range of differently modelled data, and
   (b) Suitable tools which allow efficient aggregation and presentation of data to the UI from multiple datasets. The efficient and scalable retrieval of resources is particularly important for UIs which change views and require frequent querying of various datasets. Some approaches to tackling this problem were described in [11].
2. In the real world, the data publishers and consumers may be different entities, and this is what we enforced in our case study. Our data consumers had partial knowledge of the domain and found it difficult to understand the domain and the data being modelled. This is best illustrated in the case of the hierarchy of the administrative geography. Some constituencies can be mapped into the administrative region of counties, while some are parts of

counties. Querying or browsing the data did not help in this instance. This points out the need for a mechanism, or a toolset, that helps developers give better description of the domain being modelled,

We have re-published the data we generated into the linked data cloud[16]. Resolving data and schema heterogeneity is a heuristic semi-automatic process. In future work, we aim to explore the application of data mining techniques to reduce the time it takes a human expert to align instances and/or schema. We have built a backlinking service[17] to the Linking Open Data cloud. We aim to further integrate the backlinking service to our datasets. In addition, we aim to provide an efficient scalable user interface able to visualise and search multiple datasets.

5 Acknowledgements

This work was supported by the EnAKTing project, funded by EPSRC project number EP/G008493/1. We thank Ian Millard, Hugh Glaser, and John Darlington for their contributions.

References

1. Tim Berners-Lee and Nigel Shadbolt: *Put in your postcode, out comes the data.* in "The Times" Nov 18, 2009 (Available from: http://www.timesonline.co.uk/tol/comment/columnists/guest_contributors/article6920761.ece. Last accessed 13 Dec. 2009)
2. Alani Harith, Dupplaw David, Sheridan John, O'Hara Kieron, Darlington John, Shadbolt Nigel, and Tullo Carol: *"Unlocking the Potential of Public Sector Information with Semantic Web Technology"*. pub. Lecture Notes in Computer Science, vol. 4825/2008, 2007
3. Ordnance Survey Data: http://data.ordnancesurvey.co.uk/
4. S. Harris, N. Lamb, and N. Shadbolt: *4store: The Design and Implementation of a Clustered RDF Store.* In The 5th International Workshop on Scalable Semantic Web Knowledge Base Systems (SSWS2009).
5. M. Hausenblas, W. Halb, Y. Raimond, L. Feigenbaum, and D. Ayers: *SCOVO: Using Statistics on the Web of Data.* In European Semantic Web Conference 2009 (ESWC 2009).
6. T. Tiropanis, H. Davis, D. Millard, M. Weal, S. White, and G. Wills: *Semantic Technologies for Learning and Teaching in the Web 2.0 era - A survey.* In WebSci'09: Society On-Line, 2009.
7. G. Kobilarov, T. Scott, Y. Raimond, S. Oliver, C. Sizemore, M. Smethurst, C. Bizer, and R.Lee: *Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections.* In 6th European Semantic Web Conference, ESWC 2009, pp. 723-737. pub. Springer, 2009.
8. O. Hassanzadeh and M. Consens: *Linked Movie Data Base.* In Linked Data on the Web (LDOW2009).

[16] mortality.psi.enakting.org, nhs.psi.enakting.org, crime.psi.enakting.org

[17] backlinks.psi.enakting.org

9. C.Bizer, R. Cyganiak, and T. Heath: *How to Publish Linked Data on the Web.* http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/
10. M. Hausenblas: *Linked Data Applications - The Genesis And The Challenges of Using Linked Data On The Web.* DERI TECHNICAL REPORT 2009-07-26, July 2009.
11. D. Smith and m. schraefel: *Interactively using Semantic Web knowledge: Creating scalable abstractions with FacetOntology.* Unpublished. http://eprints.ecs.soton.ac.uk/17054/ (*Last accessed 2009-12-19*).
12. Y. Kalfoglou and M. Schorlemmer: *Ontology mapping: the state of the art.* In The Knowledge Engineering Review Journal, vol. 18, 2003.
13. M. Salvadores, G. Correndo, B. Rodriguez-Castro, N. Gibbins, J. Darlington, and N. Shadbolt: *LinksB2N: Automatic Data Integration for the Semantic Web.* In International Conference on Ontologies, DataBases, and Applications of Semantics (ODBASE 2009).