Put in Your Postcode, Out Comes the Data: A Case Study

Tope Omitola[1], Christos L. Koumenides[1], Igor O. Popov[1], Yang Yang[1], Manuel Salvadores[1], Martin Szomszor[2], Tim Berners-Lee[1], Nicholas Gibbins[1], Wendy Hall[1], mc schraefel[1], and Nigel Shadbolt[1]

[1] Intelligence, Agents, Multimedia (IAM) Group
School of Electronics and Computer Science
University of Southampton, UK
{t.omitola,clk1v07,ip2g09,yy1402,ms8,timbl,
nmg,wh,mc,nrs}@ecs.soton.ac.uk
[2] City eHealth Research Centre
City University London, UK
martin.szomszor.1@city.ac.uk

Abstract. A single datum or a set of a categorical data has little value on its own. Combinations of disparate sets of data increase the value of those data sets and helps to discover interesting patterns or relationships, facilitating the construction of new applications and services. In this paper, we describe an implementation of using open geographical data as a core set of "join point"(s) to mesh 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.

Keywords: Public Sector Information, Linked Data, Public Open Data, Data Publication, Data Consumption, Semantic Web.

1 Introduction

In the private consumption of goods and services, there is a drive to put the consumer at the centre of production, i.e. to co-produce the goods and services consumers want. This drive has also come to the production and consumption of public goods and services. Citizens want governments to improve the delivery of public services. One way to improve the delivery is to put the citizen as the co-producer of these goods and services. There are concrete cases of these, where citizens are at the heart of service designs and accountability. A good example is Patients Know Best[1], where the principle is the "empowerment of health workers

[1] http://www.patientsknowbest.com/

L. Aroyo et al. (Eds.): ESWC 2010, Part I, LNCS 6088, pp. 318–332, 2010.

and patients to be creative with information". An example of co-production here is the opening up of access and communications to:

- patients' records,
- their medical doctors and consultants, and
- other information, such as treatments, care, etc.

For this to be more seamless, citizens need to have access to public data, and there will need to be linkage of these data and records across organisational boundaries to help patients make informed decisions, and to enable the quality of care to be monitored. Therefore, the provision and opening up of government data contributes to the improvement of the delivery of public services.

With a view to this, governments around the world are opening up government data to enable, among other things, this process of citizens' co-production. Questions such as: "Where can I find a good school, hospital, investment advisor, employer?", and more complex questions, such as the integration of these domains and their potential relationships, can be more easily answered than they are at present, using open government data.

In the United States, the government has set up *data.gov*[2] for the publication of public data. The United Kingdom government has set up a public data store[3] where large quantities of public sector information have been freed up or published, ranging from geospatial, statistical, financial, and legal data. While most of these data are published in spreadsheet or comma-separated-values (csv) formats, publishing them in structured machine-processable formats will be highly useful for ease of linking to other data sources and for ease of re-use.

Although the Semantic Web offers solutions to publishing data in a structured machine-processable format, the requirement for agreed ontologies has often presented a hurdle in the deployment of these technologies [2]. The Linked Data[4] movement advocates a bottom-up approach to ontology agreement [6] by publishing data in a structured machine-processable formats, such as RDF, before agreeing on ontologies for specific applications. The benefits for making data available in a structured and machine-processable formats include:

1. End users know what to expect, and find content easier to read,
2. Online services can aggregate accurate, up to date, and comprehensive information, and
3. Data re-users will find it easier to create new combinations of data that are more relevant to end users.

There have been many efforts in publishing data as linked data. Some noteworthy examples include

[2] `http://www.data.gov`
[3] `http://www.data.gov.uk`
[4] The term "Linked Data" refers to a style of publishing and interlinking structured data on the Web [`http://www.w3.org/DesignIssues/LinkedData.html`].

1. The London Gazette[5], where linked data was used to maximise the re-use of the information within it, and as a vehicle for the government to serve semantically enabled official information, and
2. The New York Times[6], which, in 2009, published its news vocabularies as linked open data.

There are some examples of linking data sources from different domains. Such examples include:

1. BBC's Music Beta[7]. This links data across BBC[7] domains with data from MusicBrainz[8], and DBpedia[9]. The BBC maintains data of programmes, music, artists, etc, but most of these information are stored in data silos. For many years, it has been difficult for the BBC to have an "across domain" look at its data; Music Beta provides a solution to this.
2. Linked Movie Database[8]. This is an RDF data source of movies, which are the results of interlinking data from disparate data sources, such as RDF Book Mashup[10], MusicBrainz, Geonames[11], and DBpedia.

A noteworthy example in the public sector is the Postcode paper[12] which, by using London (United Kingdom) postcodes, provides an integrated view of local services, environmental information, and crime statistics of neighbourhoods. However, there are challenges with providing an integrated view of linked datasets. These include:

1. (the case of) sourcing, or discovering, the appropriate datasets,
2. their formats,
3. which "join point"(s) to use,
4. selecting the normal form to use and its representations, e.g. will it be RDF/XML or RDF/N3,
5. data interlinking issues (i.e. setting links between entities in different data sources), and
6. building the client applications to consume the data

This paper describes the results and analyses from experiments carried out from selecting and combining some United Kingdom public datasets that can be used to infer relationships amongst these data. Berners-Lee and Shadbolt [1] laid out the benefits of publishing non-personal public computer-readable data for reuse. We set out to implement some of the ideas set out in their paper, and to use administrative entities (geographic) data, from the UK's Ordnance Survey, as a "join point" to mesh data for crime, mortality rates, and hospital waiting times.

[5] http://www.london-gazette.co.uk/
[6] http://data.nytimes.com/
[7] http://www.bbc.co.uk
[8] http://musicbrainz.org/
[9] http://dbpedia.org/About
[10] http://www4.wiwiss.fu-berlin.de/bizer/bookmashup/
[11] http://www.geonames.org/
[12] http://blog.newspaperclub.co.uk/2009/10/16/data-gov-uk-newspaper/

By **meshing**, we mean the ability to naturally merge together a dynamic set of information sources, and a **join-point** is a point of reference shared by all datasets. As most of these data were not linked data and were published by different departments of government, we describe the processes we undertook to convert them into a linked data format, the challenges encountered, and the solutions we devised. Building clients to consume these data also posed its own challenges, which we describe.

2 Related Work

The provision and consumption of public data have been an ongoing effort for many years. These vary from the solutions used in the linked and the non-linked data space.

2.1 Non-linked Data Space

1. **Mashups.** used for many years as a mechanism to provide and consume public data. They fuse data from two or more Web applications to create a new service. The website, *GeoWorldBank*[13] is an example of a data mashup from Google[14] and the World Bank to display a country's statistics, such as its population and its income level. A mashup application is comprised of three different components that are disjoint, both logically and physically. These are: *the API/content providers*, *the mashup site*, and *the client's Web browser*. Mashups provide many benefits, such as the provision of access to massive amounts of content which no individual could gather on their own, and lowering the barriers to developing novel applications. They have some disadvantages, and these include:

   (a) Data pollution. Some of these data sources may contain erroneous or inconsistent data, compromising the value provided by the mashup.
   (b) If a mashup uses screen scraping as part of building it, one small change in the data markup of any of the source sites may compromise the quality of the mashup.
   (c) Mashups are implemented against a fixed number of data sources and can not take advantage of new data sources that appear on the Web.

2. **Other Third-party Datastores.** There are a few charity or non-profit based institutions who provide, through APIs and bulk file downloads, data of members of a country's governmental institutions. Some are: (a) *GovTrack.us*[15], a website providing a set of tools used to research and track the activities in the U.S. Congress, using congressional district maps, or searching by name, or using the U.S. ZIP code. It provides two main access to its

[13] `http://geo.worldbank.org/`
[14] `http://maps.google.co.uk/`
[15] GovTrack.us.

data, via raw XML files and APIs, and (b) *TheyWorkForYou*[16] a website that provides data on the Members' and the Houses of the United Kingdom Parliament, through using a set of APIs and raw XML files.

2.2 Linked Data Space

There are many initiatives around the world making public data available. These include:

- DBpedia: A community effort to extract structured information from Wikipedia, linking this information to other datasets, and making it available on the Web.
- GeoNames: GeoNames integrates geographical data, such as place names, population, etc, from various sources. It gives access to users to manually edit, correct, and add new names.
- Data.gov.uk[17]: This is U.K. government central place to publish public data as RDF.
- The Ordnance Survey[18]: Great Britain's national mapping agency, providing geographic data used by government, businesses, and individuals. It provides linked data format of the administrative and voting regions in Great Britain. This data includes the names, census code, and area of the regions of Great Britain.

In DBpedia and GeoNames, as they are user-community based initiatives, there are risks of introducing data redundancy and/or data pollution. Data.gov.uk is only open to registered members, and many of the data it points to are not yet published in RDF format. The Ordnance Survey, however, is an authoritative source of the data it produces.

3 Application Case Study

3.1 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[19]. 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 have produced a number of ontologies and an RDF data set that represents the

[16] `http://www.theyworkforyou.com/`
[17] `http://data.hmg.gov.uk/about`
[18] `http://www.ordnancesurvey.co.uk/oswebsite/`
[19] `http://linkeddata.org`

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.

3.2 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 and HTML formats, while some were in XLS formats. We decided against scraping, for reasons outlined in section 2.1. 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[20]. 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[21].
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. For example, one of the datasets, the datasets for recorded crime for England and Wales 2008/09, has dimensions such the "Police Force Areas of England and Wales" and the types of recorded crimes, e.g. burglary. 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

[20] `http://data.hmg.gov.uk/about`

[21] 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`)

geographical boundaries, we used `owl:sameAs` to assert equivalences between them.

7. Consumption of data. We used Exhibit[22] 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.

4 Public Sector DataSets - Publication and Consumption

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

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

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

4.2 Modelling the Datasets

Most of our vocabularies come from Friend-of-a-Friend (FOAF)[23], Dublin Core (DC)[24], and SCOVO, thereby following the advice given in [9] to re-use terms from well-known vocabularies.

Modelling the datasets for Recorded Crimes. Each row, of the spreadsheet provided by the Home Office[25], consisted of data of each Police Force area in England and Wales, and for a particular row, its columns contained the recorded

[22] `http://simile.mit.edu/wiki/Exhibit/API`
[23] `http://xmlns.com/foaf/spec/`
[24] `http://dublincore.org/documents/dcmes-xml/`
[25] `http://www.homeoffice.gov.uk/about-us/publications/non-personal-data/`

crime values for the following offences, "Violence against the person", "Robbery", "Burglary", "Offences against vehicles", etc. The time period under discussion, i.e. 2008/09, the geographical areas and regions, and the different types of criminal offences are `scovo:Dimension`(s). A snippet of the RDF/Turtle schema representation is shown below:

```
:TimePeriod rdf:type owl:Class;
                              rdfs:subClassOf scovo:Dimension.
:TP2008_09 rdf:type :TimePeriod.
:GeographicalRegion rdfs:subClassOf scovo:Dimension;
    dc:title "Police force area, English region and Wales".
:CriminalOffenceType rdf:type owl:Class;
                              rdfs:subClassOf scovo:Dimension.
```

Modelling the Hospital Waiting List. Each row, of this dataset[26], 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".
: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[27]. 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.

[26] `http://www.performance.doh.gov.uk/waitingtimes/index.htm`
[27] `http://ukparse.kforge.net/parlparse/`

4.3 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[28] to convert the Parliament data, into RDF triples. These triples were stored in our local 4store system.

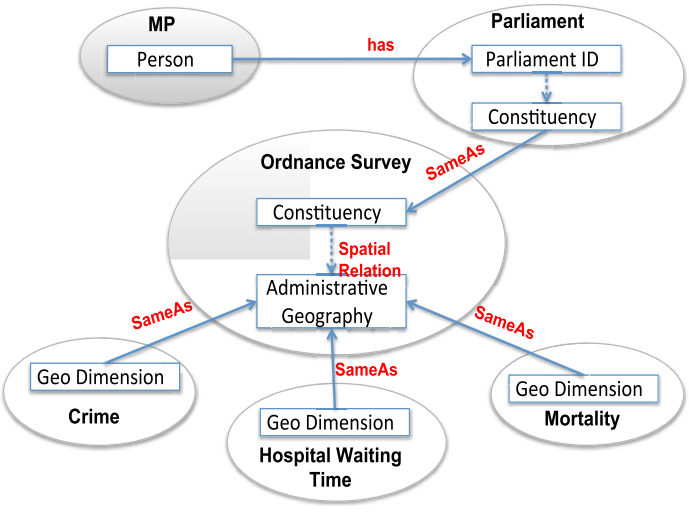


Fig. 1. Alignment of datasets using the Ordnance Survey Administrative Geography

4.4 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) (figure 1). The relevant entities here were constituencies data from the Ordnance Survey (OS).

[28] `http://jena.sourceforge.net/`

Aligning the datasets for Recorded Crimes We made use of our local 4store and also of the OS sparql endpoint at "http://api.talis.com/stores/ordnance-survey/services/sparql". Our goal was to set `owl:sameAs` equivalences between the geographical entities (geo names) in our dataset and respective (Westminster) constituencies as defined by the OS. The geographical entities in the OS are of several types, some are civil parishes, some are constituencies, and some are counties. A county usually contains a set of constituencies (and parishes). We queried the OS for the type of geo name in our dataset. If a constituency, we set the geo name to be the same as the OS identifier for that constituency. If the geo name is a county, we queried the OS for the list of constituencies it contained, and we set the geo name to be the same as the individual members' identifiers of that list. For example, the geo name Cumbria, represented in our 4store as `http://enakting.ecs.soton.ac.uk/statistics/data/Cumbria`, has many parishes, but only four constituencies. We set Cumbria to be the same as these four constituencies, shown below (to save space, only one is shown):

```
<http://enakting.ecs.soton.ac.uk/statistics/data/Cumbria>
    <http://www.w3.org/2002/07/owl#sameAs>
        <http://data.ordnancesurvey.co.uk/id/7000000000024876>.
```

However, we encountered a few special cases. A police force region in our dataset was known as "Yorkshire and the Humber Region", but the OS did not have this but had an entity called "Yorkshire & the Humber" which were the same. We manually had to query the OS for this and also to give us the constituencies under 'Yorkshire & the Humber", and set them to be the same as the "Yorkshire and the Humber Region" in our dataset.

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[29] 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[30]. 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.

[29] `http://code.google.com/apis/maps/`

[30] "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`].

4.5 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. Following on the methodology as described in the Postcode paper, 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.

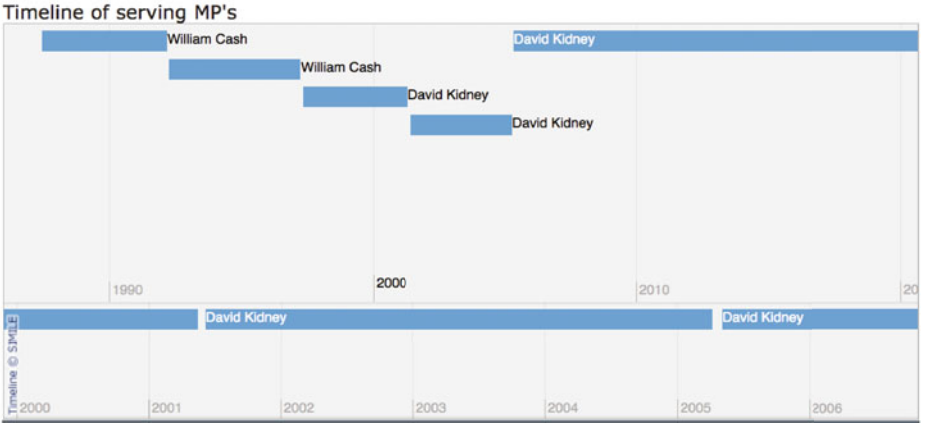
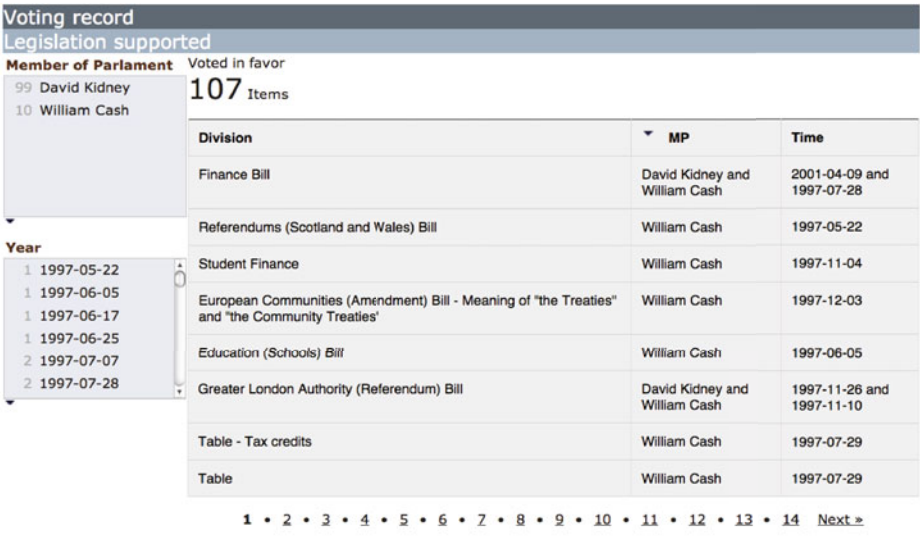


Fig. 2. Displaying data, as timeline, of some British Members of Parliament

In the application[31], 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 (Figure 2), and shows data about their voting records (Figure 3), and expenses (Figure 4). 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 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

[31] The application is at `http://psiusecase.enakting.org/`



Voting record

Legislation supported

Member of Parlament

99 David Kidney
10 William Cash

Voted in favor

107 Items

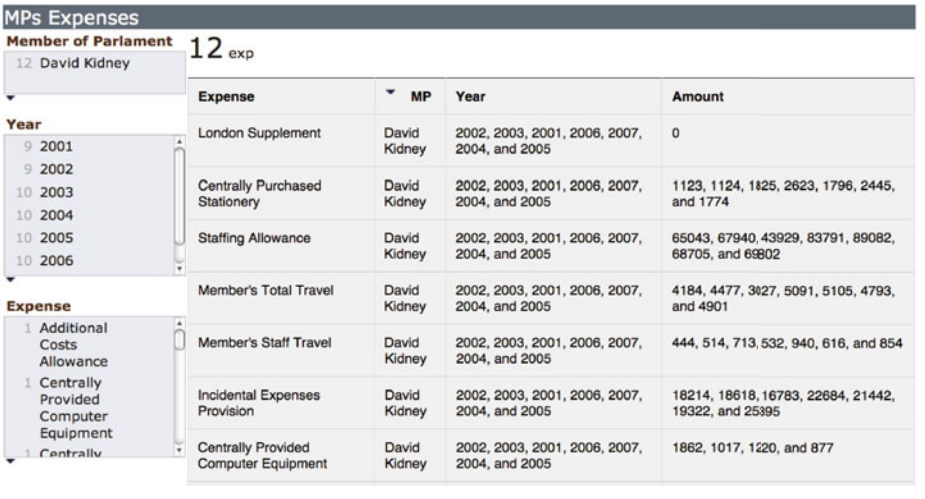
Year

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

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. 3. Displaying voting record data for selected British Members of Parliament



MPs Expenses

Member of Parlament

12 David Kidney

12 exp

Year

9 2001
9 2002
10 2003
10 2004
10 2005
10 2006

Expense

1 Additional Costs Allowance
1 Centrally Provided Computer Equipment
1 Centrally

Expense	MP	Year	Amount
London Supplement	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	0
Centrally Purchased Stationery	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	1123, 1124, 1825, 2623, 1796, 2445, and 1774
Staffing Allowance	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	65043, 67940, 43929, 83791, 89082, 68705, and 69802
Member's Total Travel	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	4184, 4477, 3027, 5091, 5105, 4793, and 4901
Member's Staff Travel	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	444, 514, 713, 532, 940, 616, and 854
Incidental Expenses Provision	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	18214, 18618, 16783, 22684, 21442, 19322, and 25395
Centrally Provided Computer Equipment	David Kidney	2002, 2003, 2001, 2006, 2007, 2004, and 2005	1862, 1017, 1220, and 877

Fig. 4. Displaying expenses data for a selected British Member of Parliament

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.

5 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.

5.1 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.

5.2 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[32]. 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[33] 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.

Acknowledgements

This work was supported by the EnAKTing project, funded by EPSRC project number EI/G008493/1.

References

1. Berners-Lee, T., Shadbolt, N.: Put in your postcode, out comes the data. In: The Times (November 18, 2009),
`http://www.timesonline.co.uk/tol/comment/columnists/guest_contributors/article6920761.ece`
(last accessed December 13, 2009)

[32] `mortality.psi.enakting.org`, `nhs.psi.enakting.org`, `crime.psi.enakting.org`
[33] `backlinks.psi.enakting.org`

2. Harith, A., David, D., John, S., Kieron, O., John, D., Nigel, S., Carol, T.: Unlocking the Potential of Public Sector Information with Semantic Web Technology. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L.J.B., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 708–721. Springer, Heidelberg (2007)
3. Ordnance Survey Data, `http://data.ordnancesurvey.co.uk/`
4. Harris, S., Lamb, N., Shadbolt, N.: 4store: The Design and Implementation of a Clustered RDF Store. In: The 5th International Workshop on Scalable Semantic Web Knowledge Base Systems SSWS 2009 (2009)
5. Hausenblas, M., Halb, W., Raimond, Y., Feigenbaum, L., Ayers, D.: SCOVO: Using Statistics on the Web of Data. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 708–722. Springer, Heidelberg (2009)
6. Tiropanis, T., Davis, H., Millard, D., Weal, M., White, S., Wills, G.: Semantic Technologies for Learning and Teaching in the Web 2.0 era - A survey. In: WebSci'09: Society On-Line (2009)
7. Kobilarov, G., Scott, T., Raimond, Y., Oliver, S., Sizemore, C., Smethurst, M., Bizer, C., Lee, R.: Media Meets Semantic Web - How the BBC Uses DBpedia and Linked Data to Make Connections. In: Aroyo, L., Traverso, P., Ciravegna, F., Cimiano, P., Heath, T., Hyvönen, E., Mizoguchi, R., Oren, E., Sabou, M., Simperl, E. (eds.) ESWC 2009. LNCS, vol. 5554, pp. 723–737. Springer, Heidelberg (2009)
8. Hassanzadeh, O., Consens, M.: Linked Movie Data Base. In: Linked Data on the Web, LDOW 2009 (2009)
9. Bizer, C., Cyganiak, R., Heath, T.: How to Publish Linked Data on the Web, `http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/`
10. Hausenblas, M.: Linked Data Applications - The Genesis And The Challenges of Using Linked Data On The Web. Deri Technical Report 2009-07-26 (July 2009)
11. Smith, D., Schraefel, M.: Interactively using Semantic Web knowledge: Creating scalable abstractions with FacetOntology (unpublished), `http://eprints.ecs.soton.ac.uk/17054/` (last accessed 2009-12-19)
12. Kalfoglou, Y., Schorlemmer, M.: Ontology mapping: the state of the art. The Knowledge Engineering Review Journal 18 (2003)
13. Salvadores, M., Correndo, G., Rodriguez-Castro, B., Gibbins, N., Darlington, J., Shadbolt, N.: LinksB2N: Automatic Data Integration for the Semantic Web. In: International Conference on Ontologies, DataBases and Applications of Semantics, ODBASE 2009 (2009)