Unlocking the Potential of Public Sector Information with Semantic Web Technology

Harith Alani[1], David Dupplaw[1], John Sheridan[2], Kieron O'Hara[1], John Darlington[1], Nigel Shadbolt[1], and Carol Tullo[2]

[1] Intelligence, Agents, Multimedia
School of Electronics and Computer Science
University of Southampton, Southampton, UK
{h.alani, dpd, jd, kmo, nrs}@ecs.soton.ac.uk
[2] Office Of Public Sector Information
Admiralty Arch, The Mall
London, UK
{john.sheridan, carol.tullo}@opsi.x.gsi.gov.uk

Abstract. Governments often hold very rich data and whilst much of this information is published and available for re-use by others, it is often trapped by poor data structures, locked up in legacy data formats or in fragmented databases. One of the great benefits that Semantic Web (SW) technology offers is facilitating the large scale integration and sharing of distributed data sources. At the heart of information policy in the UK, the Office of Public Sector Information (OPSI) is the part of the UK government charged with enabling the greater re-use of public sector information. This paper describes the actions, findings, and lessons learnt from a pilot study, involving several parts of government and the public sector. The aim was to show to government how they can adopt SW technology for the dissemination, sharing and use of its data.

1 Introduction

Public Sector Information (PSI) can make an important contribution to bootstrapping the SW, which in turn will yield many gains. UK government tends to see the web primarily as a medium for the delivery of documents and the dissemination of content to the citizen. With the emergence of a re-use policy agenda for PSI, the UK government is beginning to develop a far richer and deeper understanding of the SW and the contribution it can make in terms of achieving greater efficiency through information sharing and integration to realise broader economic and social gains.

The Office of Public Sector Information (OPSI)[3] is responsible for the management of all of the UK government's intellectual property, including setting standards, delivering access and encouraging the re-use of PSI. In the UK, any work produced by an employee of the government is deemed to be owned by the Crown and thus subject to Crown copyright. Under this constitutional position, Carol Tullo, the Director of OPSI and co-author to this paper, is granted authority by Her Majesty The Queen to manage

[3] Developed from Her Majesty's Stationery Office (HMSO)

all copyrights and databases owned by the Crown. OPSI also has an important role as a regulator of holders of public sector information (e.g. the Met Office, Ordnance Survey) for their information trading activities.

In the UK, large quantities of public sector information have been made available, ranging from geospatial, statistical, financial and legal information. However, making data available and making data reusable are two very different things. Most government data is published online in text formats with little structure, thus inhibiting its re-use. By using unstructured, non-semantic representations of the data, it becomes almost impossible for machines to find or understand and integrate this rich source of information. For these reasons, OPSI decided to initiate a research project, AKTivePSI. The aim of AKTivePSI is to show how the use of SW technology can facilitate the large scale integration and re-use of public sector information, ultimately to the benefit of government, business and citizen alike. AKTivePSI was about building prototypes and demonstrators to mainly win the hearts and minds of some government agencies and show them how, and what will it take, to become semantically enabled. Several of the organisations that actively participated are now investing in SW technologies, as will be highlighted in section 6.

In the following sections we will report on the decisions, actions, and results of AKTivePSI, which involved several government and information trading organisations that collect, store, and publish public sector information.

2 Related Work

The UK has developed a strong e-Government agenda over the last ten years, initially focussed on providing access to information and more latterly on delivering public services online. The publication of the government's IT strategy document, "Transformational Government - Enabled by Technology" [1] in 2005 marked an important shift in the UK government's thinking to a much broader technology agenda.

Crucially the government has identified overcoming problems with information sharing as being integral to transforming services and reducing administrative burdens on citizens and business. The UK government is committed to leveraging and producing open standards, and the GovTalk programme[4] has key documents that describe interoperability frameworks and metadata standards. With this in place, the scene is ideally set for SW technologies now to take centre stage. To use the new parlance, transformational government will require the use of transformational technology for information sharing.

The Access-eGov [8] project has been investigating how current governmental websites may be annotated using a shared reference ontology and intend to roll out methodologies on a test-bed of Eastern European governmental websites. They suggest that guided markup of current web-pages and content is perhaps the way to go. However, they correctly write that developers do not have the necessary domain knowledge to create the reference ontology, creating an extra layer of bureaucracy in the development of the system [9]. Similarly the Quebec government in Canada have embarked

[4] `http://www.govtalk.gov.uk/`

on producing a SW-service-based portal, also using a reference ontology to markup the government's web-pages.

The BRITE [18] project is building a SW infrastructure for specific areas of governmental record keeping, in this case European-wide business registrations. Vitvar and colleagues [19] explain how SW services can be used as part of the proposed Pan-European E-Government Services (PEGS) proposal, which will go some way to addressing the follow-on problem of how to integrate semantic data from different countries.

Information integration is of great importance in B2B scenarios. There are several advantages in using ontology-based architectures for information integration, such as ease of mapping, handling of different terminologies, explicit data models, etc. [3].

Using Semantic Web Services (SWS) for the integration and sharing of distributed data sources has also been suggested and demonstrated in B2B scenarios [13, 4]. Existing B2B standards for data exchange usually require considerable effort from organisations to agree how exactly they are to be used and implement that [13]. SWS is offered as an alternative to describe and discover information. This approach could allow for dynamic integration of resources, assuming that they have been appropriately described in a SWS language (eg WSML, WSDL).

3 AKTivePSI

Information policy has developed quite quickly in the UK over the last five years, with Freedom of Information legislation as well as the EU Directive, but no large scale work had been done to research the potential for reuse using SW technologies and approaches. OPSI initiated AKTivePSI as an exemplar to show what could be achieved if public sector information was made available for reuse in an enabling way.

3.1 Aims of AKTivePSI

Integrating and sharing information from distributed sources contains several obstacles and problems [5], such as scalability, different terminologies and formats, cost, etc. After meeting with the AKTivePSI government participants, we noticed that many of them shared the following misguided opinions or beliefs:

- Ontologies are very large, complex, and expensive data models
- Everyone has to agree and adopt the same terminology to enable data sharing
- To participate in the SW, their existing data infrastructures will need to be replaced with new technology
- Opening access to data only benefits the consumer, and not the provider

Our first task in this project was to correct the above misunderstandings to gain the support of the participants and encourage to provide data and some resources. The initial aims of the project were to draw together a sufficiently large set of heterogeneous information from a selection of public sector organisations in order to explore: (a) How SW technology can help turn government information into re-useable knowledge to fuel e-government, (b) Investigate the best practical approach to achieve this goal, in terms

of collecting data and constructing ontologies (c) Show how can data be integrated, and identify existing government taxonomies that are useful for this task, and (d) provide evidence that there is added value from undergoing this process.

Throughout the project, we had regular consultations with many government organisations, including the London Boroughs of Camden[5] and Lewisham[6], Ordnance Survey[7] (OS), The Stationary Office[8] (TSO), The Met Office[9], The Environment Agency[10], The Office of National Statistics[11] (ONS), and several others.

To help focus the requests for data, information was collected from the geographical area covered by two of the participating London local authorities; Camden and Lewisham.

3.2 Design Decisions

The AKTivePSI project set out to deal with real data, plenty of it, and several, very busy, data providers, keen to find solutions to their knowledge problems. In such *real world* scenarios, it becomes vital to follow a realistic approach that is practical and inexpensive. To this end, the following decisions were made at the start of the project which turned out to have a very positive impact on the project as a whole:

- No disruption to the participants' existing data flows and models. A complete and sudden transition to semantic knowledge bases (KB) is unnecessary and impractical in the short term.
- Minimum cost to the participants. They provide the data, and we provide everything else (ontologies, KB infrastructure, tools for integration, etc.). Data to be delivered in any shape, format, and delivery method. No data preparation is required from the provider. Aim here is to encourage participation, and once the benefits of the SW become more apparent, they will be more willing to invest in this new technology. The outcomes of this project show that this approach has paid off very well.
- Simulate a real-life scenario. In other words, what we build and do can be done the same way outside our lab environment. For example, we treat the KBs as if hosted by the participants.
- Small, well focussed ontologies. It is not realistic to assume that an organisation will build one monolithic ontology for all their data, or that different organisations will agree on one semantic model. Therefore, a new ontology will be constructed for each dataset, and will be designed to represent *only* the data stored in this database, rather that the extended domains that the data might be related to (examples later). These numerous, small ontologies will be mapped together to form a small SW.

[5] `http://www.camden.gov.uk/`
[6] `http://www.lewisham.gov.uk/`
[7] `http://www.ordnancesurvey.co.uk/`
[8] `http://www.tso.co.uk/`
[9] `http://www.metoffice.gov.uk/`
[10] `http://www.environment-agency.gov.uk/`
[11] `http://www.statistics.gov.uk/`

– Data provenance must be preserved. Each dataset provided to us was transferred into a *separate* KB with its own ontology to eliminate any risks of data contamination from one database to another. Furthermore, each ontology contains a few classes and properties to represent the source of data, including name of supplier, name of data set, date supplied, etc. Source information is also attached to all triples when stored in the triple store.

4 Public Sector Datasets

Several organisations who participated in AKTivePSI made some of their databases available for the project. The data was provided in various formats, including Microsoft SQL databases, Microsoft Excel Spreadsheets, text-dumps from databases, XML files, and Microsoft Access spreadsheets. We developed a number of scripts to automatically convert this data to RDF, in correspondence with their designated ontologies. Table 1 lists the data sets that we used in this work, the number of RDF statements generated for each, and a brief description of the data.

Camden Borough Council			
Land and Property Gazetteer	2.3M	Excel	Properties in Camden, full address, coordinates, type (residential/non-residential/mixed).
Food Premises	84K	Excel	Food related premises in Camden, their business names, hygiene inspection results, addresses, (eg restaurant, school, bar).
Local Businesses	170K	Excel	Businesses in Camden, names, addresses, contact info, and type of business.
Licences	100K	MSSQL	Licences for businesses in Camden, their addresses, licence types, and expiry dates.
Councillors and Committees	29K	Excel	Councillors and committees, sub committees, who sits on which committee, councillor's personal information.
Meeting Minutes	106K	Text	Web pages of committee's meeting minutes.
Lewisham Borough Council			
Land and Property Gazetteer	4M	Excel	Properties in Lewisham, their full addresses, and coordinates.
Property Tax Bands	10K	Excel	Tax property references, description, rate payers, rate value, and a one string addresses.
Ordnance Survey (data for Camden and Lewisham only)			
Address Layer 1	768K	XML	Data about buildings, addresses, and coordinates.
Address Layer 2	11.7M	XML	Data about buildings, addresses, and coordinates and building classifications (e.g. hospital, university).
PointX POI	467K	XML	Various landmarks and businesses, with names, addresses, and coordinates.
The Stationery Office London Gazette (entire database was provided, but only the below was used)			
Administration Notices	120K	Text	Notices for the appointment of administrator for corporate insolvencies.
Deceased Estates	3.2M	Text	Decease notices of individuals, names, addresses, description and date of death, address of representatives.

Table 1. Datasets provided to AKTivePSI, the number of RDF triples we generated for each dataset, and a description of what the data is about

Once we receive a new database, we (1) design and build an ontology for this data, (2) convert the data to RDF triples and store in a triple store, and (3) map the data and ontology to our existing ontologies and KBs. These stages are described in the following sections.

4.1 Ontology Construction

Ontologies vary according to their formality levels, the purpose for which they are built, and the subject matter they represent [15]. One of the recommended first steps towards building an ontology is to scope its domain to make sure the ontology does not grow too large for what is needed [14][16].

The appropriate size for an ontology depends on its purpose and the domain it represents. Some ontologies are designed to represent entire domains, and thus tend to be of very large sizes, such as the Gene Ontology (GO) [12], and Foundational Model of Anatomy (FMA) ontology [13]. Ontologies may also be built to serve the needs of specific applications and thus their sizes, though dependent on the needs of these applications, tend to be much smaller than the domain encapsulating ontologies. Other ontologies, as in our case, are data-dependent, where they are mainly built to represent a collection of data, to improve accessibility and understandability of the data. The scale of such dataset-specific ontologies is limited to the scope of the data.

As stated earlier, one of our principals for this project was to ensure the ontologies we build for the provided datasets are of low complexity and limited in scope and size. Small ontologies are cheaper and easier to build, maintain, understand, and use. In AKTivePSI, we found that most of the databases held by the participating organisation only required a small number of concepts and relationships to represent the stored data.

In AKTivePSI, we wanted to show that ontologies are not hard to build if limited to representing databases of defined scopes. We also wanted to show that it is not necessary to come to a common, agreed concensus on vocabulary, but that through ontology mapping techniques, locally-built ontologies can also prove very useful. Figure 1 shows an example of a ontology we have built, that describes, in very simple terms, the domain of Camden's Land and Property Gazetteer. In total, we constructed 13 ontologies, one for each dataset listed in table 1. All the ontologies were in OWL DL, and were mainly used to control vocabulary and to cross-link knowledge bases.

4.2 Generating RDF

From the ontology we are able to create instances by running simple scripts over the data to produce RDF. The scripts were hand-rolled specifically for the database and ontology which they were linking (reused across similar databases and ontologies). Although they were manually built, a framework for semi-automatic script generation would not be inconceivable. The scripts were highly reusable and hence were very easy to tune for new datasets and ontologies. We 'demonstrated to the participants the relative ease of converting legacy data to RDF using cheap and ordinary technology.

As shown in table 1, the total number of RDF triples that we generated for the government data exceeded 23 million. So although we needed small ontologies, we also needed scalable KB to hold all these RDF triples. We used the 3Store [6], an RDF triple-store developed in the AKT project, to store the generated RDF files. This triple-store provides a SPARQL endpoint, which is a servlet that accepts SPARQL queries and returns results in XML.

[12] `http://www.geneontology.org/`
[13] `http://sig.biostr.washington.edu/projects/fm/`

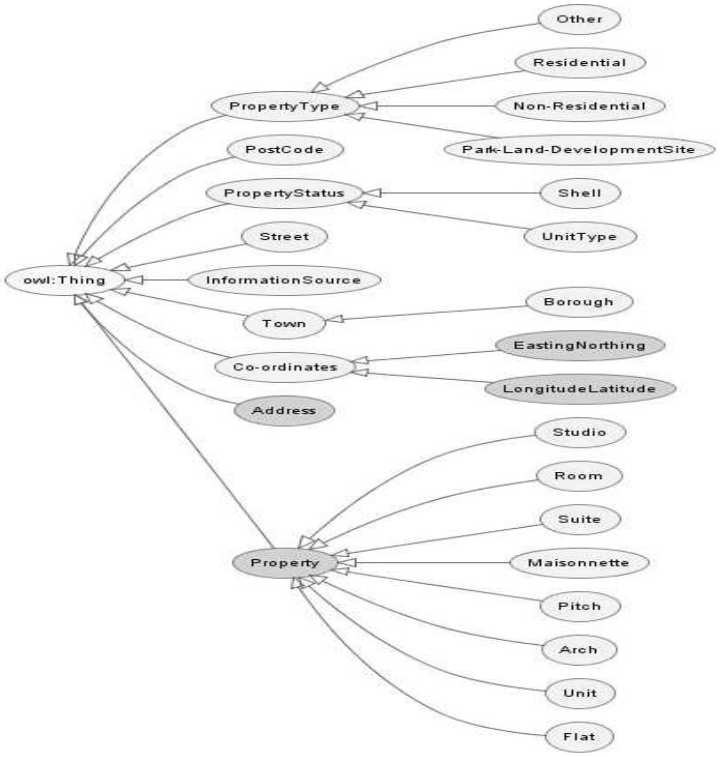

Fig. 1. Ontology for the Camden Land and Property Gazetteer

4.3 Webbing the Knowledge

One of main reasons for utilising ontologies is that the need for tight physical integrations between systems is removed [17]. Ontologies allow integration to happen using 'soft' mappings between concepts and instances that queries or data browsers can follow to find similar or duplicated entities. In our work, we used the special *owl:sameAs* property to link any mapped entities. By connecting our KBs in this way we are able to provide much greater flexibility and querying power than the original data structures could provide.

One main aim of this work is to show the added value of using SW technology for publishing and using government data. Forming a bigger semantic network by integrating the KBs containing all the participants' data will add even more value to the data, and ease communication and data exchange between the partners.

We performed three levels of mappings:

– Mapping of local ontologies. It is safe to assume that individual organisations will know most about any of the ontologies they develop for their data, and hence it is possible for these local ontologies to be mapped to each other. For example, we developed two ontologies for datasets from Lewisham. Each ontology has

classes representing Property, Address, Post Code. These concepts were links with owl:sameAs to indicate that they represent the same concepts. Another example is mapping the concept Premises from the Food Premises ontology of Camden to the Property class in the Land and Property ontology of Camden. To semi-automate these mapping, we used CROSI [7], a freely available tool that offers a wide choice of mapping algorithms.

– Mapping of instances. Because we are using a data-centric approach, it was very useful to map the instance data to each other as well. For example the instance *post-code_N6_6DS* in one KB maps to the instance *pc_N66DS* in another. Since these instances really do refer to the same object we are able to infer much more data about certain objects that refer to this instance. In fact, we found that simply linking on one data object (the postcode) was enough to gleen useful information from various datasets to such an extent that mashups are made easier (section 4.5). Instance mappings were done automatically using simple scripts that search for duplicates of specific type of instance (e.g. postcodes, streets, councillors). An owl:sameAs link will be automatically added between the corresponding instances once such a mapping is found.
– Mapping of local ontologies to the government reference taxonomy; IPSV. IPSV (the Integrated Public Sector Vocabulary) is a "structured list of terms for the Subject metadata of public sector resources" [2]. UK e-Government Metadata Standard requires public sector organisations to comply with IPSV. AKTivePSI partners expressed some difficulties mapping their databases to IPSV, and hence part of this project was explore this taxonomy and assess its suitability for this task. To better understand the problem, we manually mapped our ontologies to the best matched terms in the IPSV.

4.4 Exploring the Knowledge Network

Now that all the data is ontologically represented and stored in KBs, we need to demonstrate to the participating government organisations what and where the added value is.

RDF provides a well-understood grounding on which data may be shared, and this in itself provides added value, such that re-use of the data is made much easier (see section 4.5 on Mash-ups).

4.5 Mashing-up Distributed KBs

Once data is available in easily parsable and understandable formats, such as RDF, mash-ups become much easier to generate by searching RDF KBs and mashing-up data on the fly, which is one of the advantages the SW promises. Two examples of such mash-ups were created in AKTivPSI. The aim of building these mash-ups was to demonstrate the relative ease with which they can be constructed from semantically represented knowledge.

The Camden Food Premises database gives information about the hygiene check results and health risk of various premises around the Camden area that handle food. The risk categories are given a level between A, which is high risk, to E which is low risk, and is based on the cleanliness of the premises, compliance with regulations, type

of preparation that is performed, etc. The Food Premises database contains lots of information on these properties, but displaying this information on a map is difficult because the geographical co-ordinates are missing from this particular data set.

However, the Ordnance Survey's Address Layer and Points of Interest (PointX) datasets contain easting and northing coordinates for businesses and properties. The instance mapping of postcodes we performed earlier helped to cut down our search space for finding matching addresses in the datasets. Indeed, once we had found matches we were able to assert them as being the same, thereby avoiding the need for searching again.

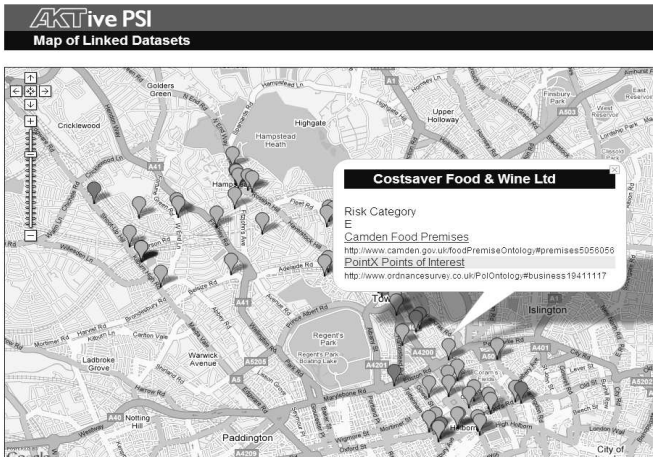

Fig. 2. Google Maps mashup of the Camden Food Premises dataset made possible by mapping the data to the OS Address Layer II and PointX dataset

To create the mash-up, a number of SPARQL queries were written that searched for each premises' address from the Food Premises dataset in each of the OS two datasets and once a match is found the co-ordinates are retrieved and the premises is displayed on a Google map. The information from Food Premises along with the mapping performed between one dataset and another, provides extra context to instances from both datasets. The PointX dataset gains access to the risk level of the food premises (as well as the implicit knowledge that the premises are used for preparing food), and the food premises dataset garnered exact coordinates for the premises. Figure 2 shows a simple Google Maps mash-up that uses the mapping to provide a visual display of the food premises dataset.

This type of mash-up could be very good for public awareness (and therefore commercial competition). For example, one particular business that scored within the high risk category, has glowing customer reviews on restaurant review sites across the internet.

As for Lewisham, we were able to use the PointX dataset for a similar use for the Lewisham Land and Property Gazetteer. This dataset contains information about all

kinds of properties across the Lewisham ward, and includes address and coordinate information; however, it does not contain information about the business inhabiting a property – information that the PointX data does provide. We provided a mash-up that shows the location of business properties

5 Findings

Introducing a new technology, such as the SW to any organisation must be managed very carefully to avoid any disruption to their current work procedures and data flow, and to gain their trust and interest in the new technology. Below are some of the findings and lessons learnt from the AKTivePSI study that relates more specifically to government agencies.

Minimise disruption to existing infrastructure: It was critical to show that adopting SW technology does not mean throwing away existing database technologies; the re-engineering of governmental information processing is a complex and difficult task that is facilitated by special conditions and structures that do not obtain in the United Kingdom [11]. An important part of our task was to show that the costs of SW adoption were relatively low. We demonstrated how simple scripts can be used to convert all their data into RDF triples. The approach we adopted in AKTivePSI was to *cache* the given databases into triple stores. However, this might not be the best solution as it duplicates existing databases. A slightly different approach is suggested in [10], where they imported the data into ontologies, then exported it back into relational databases with new structures that are closer to those of the used ontologies. In other words, they changed the database schema to match the ontology. We are now experimenting with an alternative approach, which is to use a technology like D2RQ[14] which enables layering an ontology on top of a non-RDF database, thus removing the need to duplicate or change the structure of the original database. Such technology maintains the benefits of scalability and maturity of RDBMS, as well as providing RDF and SPARQL access points.

Minimal ontological commitment: Constructing ontologies requires certain skills and modelling knowledge and expertise. Government organisations worry about the possible high cost of building such complex knowledge structures.

We were encouraged by the results of applying SW technologies to governmental data. Not only were the benefits very high, even on a small quantity of data, but the costs were relatively low. The building of small, data-centric ontologies was an easily achievable goal for developers in governmental institutions, possibly working on limited budgets. This is a lesson that is of interest in the wider Semantic Web field, as arguments continue over the overhead that developing and maintaining ontologies will require. As a result of this work, Camden Council and the London Gazette are now developing their own ontologies to represent some of their datasets.

Some of these organisations thought that Cyc, Gene Ontology, or even IPSV, are the sort of ontologies they need to build to become semantically enabled. We demonstrated a cheap and practical approach, where ontologies are scaled to individual datasets rather

[14] `http://sourceforge.net/projects/d2rq-map/`

than to entire domains, then *gradually* linked together to enable data sharing. It is possible that more elaborated ontologies might be required later on if more automation is needed for ontology mapping or for data inconsistency checking.

Extending IPSV: One of the initial concerns that some AKTivePSI participants had was the difficulty they were facing in mapping their data collections to IPSV (section 4.3). During our investigation, we found that IPSV is mainly designed to represent subject topics, not data. For example, IPSV contains more than 30 terms related to road issues (e.g. *Road safety, Road signs, Road cleaning*), but there is not a "Road" term to map a specific road to. Our conclusion was that IPSV is simply not designed to be a reference ontology for representing data and hence a different reference ontology, or an extension to IPSV, for mapping and sharing data. In AKTivePSI, we mapped each ontology to IPSV to demonstrate how IPSV can be extended to cover the required semantics.

The ability to map ontologies together provided a much more practical and less expensive alternative to agreeing or using the same terminology, which some government organisations thought was required to share data. They realised that it is possible to continue using their local terminologies whilst being able to open data exchange channels between different, distributed, databases.

Showing added value: The goal of providing better access to data is naturally not enough to win the interest, support, and active participation of data providers. It was vital to show examples of where and what is the added value of integration and shared access. Most of the organisations we met with had some needs, and sometimes laborious procedures, for acquiring data from other government sources. We illustrated the direct benefits of participating in a semantically enabled data exchange channel, especially with respect to data consistency checking, relative ease of integration and distributed querying, data exchange and merging, and lowering the cost of meeting the requests of the public for data as well as the requests of the government for providing better access to public sector information.

Data integration from multiple sources adds the value of knowledge augmentation and verification. Integrating datasets can provide useful insights into the quality of the dataset for the data provider involved. For example, the Ordnance Survey's *Address Layer 2* dataset provides a list of businesses, including their address and their geolocation, and similarly so does the PointX dataset. However, we found that the two lists of businesses do not match, where some are present in one dataset but not in the other. In some examples, the PointX dataset contained several businesses listed at the same address, while only one was listed in the OS Address Layer 2. Was this an error? Perhaps, due to the lack of temporal information, one business took over the building from another, or perhaps one business is sited in the same building on a different floor to another business. It is difficult to infer an answer, but the integration has provided some information about the quality of the datasets and made such comparisons and cross-matchings possible.

This, of course, applies equally to errors and inconsistencies in the datasets, as well as knowledge gaps. Table 2 gives an overview of one of the examples of inconsistencies in the datasets. The Sunrise Food Mart appears in the PointX database at number 354, whereas it occupies a number of building plots and is called Sunrise Food Market in the Camden Food Premises. The Ordnance Survey has an error in its naming of the

business, and says the business is at number 352, with some courier service at 354. Such inconsistencies cannot easily be automatically resolved, unless a number of other linked datasets are able to provide evidence that supports one or other of the possible addresses. All three sources are official and trusted, and hence not one can be taken as necessarily the correct one.

Name	Number	Postcode	Dataset
Sunrise Food Mart	354	NW62QJ	PointX
Sunrise Food Market	352–354	NW62QJ	Camden Food Premises
Sunrise Food Mart,352	352	NW62QJ	Ordnance Survey Address Layer 2
London No.1 Courier	354	NW62QJ	Ordnance Survey Address Layer 2

Table 2. Inconsistencies on one entity highlighted by the integration

As well as spotting many knowledge overlaps, we also identified several knowledge gaps between various participants. For example, the OS desires to get automatic feeds about accepted applications for property extensions from local councils, local councils need to receive automatic notifications from Land Registry when a property changes hands, and local councils in London would like to know when a business publishes its insolvency notice in the London Gazette. Although AKTivePSI did not implement any of these capabilities, but it showed how the SW can support such processes. Some of these services will be implemented in the second phase of AKTivePSI due to start in the coming few weeks.

Provenance and Privacy: Many agencies and institutions are instinctively secretive about their data. The SW vision is to remove human processing from the knowledge acquisition process as far as is feasible, and the idea of publishing data without even controlling the context of its presentation is of course very new in governmental circles.

At present, the *ideal* limits to data publication are unknown. A number of agencies lack understanding of what data they actually possess. These agencies needed to be assured that with SW technology, they will be able to *pick and choose* which data to share and which data to keep locked-up.

Some of AKTivePSI government participants expressed their great unease and worry about possible misuse of the data, once access and reuse are enabled with the SW. Privacy is a complex issue, with post-Enlightenment concepts under technological threat from a number of directions, not only government. Many of us are prepared to surrender our privacy for gains in efficiency or monetary benefit; others defend personal privacy as a vital pillar of a liberal democratic society. Unless and until such political dilemmas are resolved, governments will of necessity have to tread carefully when considering how far to exploit information-processing technologies such as the Semantic Web [12]. Technologies and protocols currently under development in the W3C to create a *policy-aware Web*, allowing information users, owners and subjects to express policies for information use and negotiate about them, will help make the situation clearer [20].

6 Conclusions and Future Work

The adoption of Semantic Web technology to allow for more efficient use of data in order to add value is becoming more common where efficiency and value-added are important parameters, for example in business and science. However, in the field of

government there are other parameters to be taken into account (e.g. confidentiality), and the cost/benefit analysis is more complex. The work reported here was intended to show that SW technology could be valuable in the governmental context.

An important outcome of the project is the level of awareness that has been built up in government about the potential of SW technology. Having seen what the SW technology is capable of, and the success of the pilot study of AKTivePSI, OPSI is now funding a second project which will focus on implementing and running some of the services and capabilities studied in the first stage at the premises of some of the participating government agencies.

Some of the direct outcomes of this work are: (a) the London Gazette is currently building OWL ontologies to represent parts of their data, and is working towards publishing this data in RDF; (b) OPSI oversaw the development of a URI schema, which is now being used to generate URIs for government official legislations and copyright statements; and (c) Camden Borough Council added a SW engineer to their staff force to help the council in their effort to join the SW.

AKTivePSI has given a glimpse of what is possible by applying SW technology to public sector information. We showed that by using small, purpose-built ontologies and mapping these together, greater value can be sought in the data, and re-use of the data in mash-ups becomes much easier, which should increase public awareness and access to the data.

The issue of providing better access to to public sector information has also been identified in the policy review, "The Power of Information", conducted by Prime Minister's Strategy Unit[15]. The review aims to position the UK Government in response to developments in the use and communication of citizen and state generated information on the web. The work described in this paper predates that review and helped inform the review team's analysis. Of particular note is the proposal to link the social power of the web to help address the re-usable format issue, by providing citizens with an online facility for people to come together, to discuss and formally request public sector information assets in a particular format.

The commercial re-use of public sector information that the SW enables, opens up countless opportunities for the development of new information products and services, driving forwards and accelerating the development of the knowledge economy.

References

1. Cabinet Office, Transformational Government: Enabled by Technology, Crown Copyright, Cm 6683, 2005.
2. S. D. Clarke. Guide to meta-tagging with the IPSV. Instructional, http://www.esd.org.uk/documents/ipsvhowtometatag.pdf, e-Government Unit, Cabinet Office, UK, 2005.
3. D. A. Dimitrov, J. Heflin, A. Qasem, and N. Wang. Information integration via an end-to-end distributed semantic web system. In *Proc. 5th Int. Semantic Web Conf. (ISWC)*, pages 764–777, Athens, GA, USA, 2006.

[15] `http://www.cabinetoffice.gov.uk/strategy/`

4. A. Duke, M. Richardson, S. Watkins, and M. Roberts. Towards B2B integration in telecommunications with semantic web services. In *Proc. 2nd European Semantic Web Conf. (ESWC)*, pages 710–724, Crete, Greece, 2005.
5. A. Y. Halevy, N. Ashish, D. Bitton, M. Carey, D. Draper, J. Pollock, A. Rosenthal, and V. Sikka. Enterprise information integration: Successes, challenges and controversies. In *Proc. ACM SIGMOD Int. Conf. on Management of data*, pages 778–787, New York, USA, 2005.
6. S. Harris and N. Gibbins. 3Store: Efficient bulk RDF storage. In *Proc. 1st Int. Workshop on Practical and Scalable Semantic Systems (PSSS'03)*, pages 1–20, Sanibel Island, FL, USA, 2003.
7. Y. Kalfoglou, B. Hu, D. Reynolds, and N. Shadbolt. CROSI: Capturing, Representing, and Operationalising Semantic Integration. Technical Report 11717, University of Southampton, Southampton, UK, 2005.
8. R. Klischewski. Migrating small governments' websites to the semantic web. In *In Proc. The Semantic Web meets eGovernment, AAAI 2006*, Stanford University, California, March 2006.
9. R. Klischewski and M. Jeenicke. Semantic web technologies for information management within e-government services. In *In Proc. of 37th Hawaii Conference on System Sciences 2004*, 2004.
10. A. Maier, H.-P. Schnurr, and Y. Sure. Ontology-based information integration in the automotive industry. In *Proc. 2nd Int. Semantic Web Conf.*, pages 897–912, Sanibel Island, FL, USA, 2003.
11. K. O'Hara and D. Stevens. Democracy, ideology and process re-engineering: realising the benefits of e-government in singapore. In *Proc. WWW06 Workshop on e-Government: Barriers and Opportunities*, Edinburgh, 2006.
12. K. O'Hara and D. Stevens. *inequality.com: Power, Poverty and the Digital Divide*. Oxford: Oneworld, pp.243-71, 2006.
13. C. Preist, J. Esplugas-Cuadrado, S. A. Battle, S. Grimm, and S. K. Willieams. Automated business-to-business integration of a logistics supply chain using semantic web services technology. In *Proc. 4th Int. Semantic Web Conf. (ISWC)*, pages 987–1001, Galway, Ireland, 2005.
14. D. Skuce. Conventions for reaching agreement on shared ontologies. In *Proc. 9th Banff Knowledge Acquisition for Knowledge-Based Systems Workshop*, Banff Conference Centre, Banff, Alberta, Canada, 1995.
15. M. Uschold. Building ontologies: Towards a unified methodology. In *Proc. of 16th Conf. Expert Systems*, Cambridge, UK, 1996.
16. M. Uschold and M. Gruninger. Ontologies: principles, methods and applications. *The Knowledge Engineering Review*, 11(2):93–136, 1996.
17. M. Uschold and M. Gruninger. Ontologies and semantics for seamless connectivity. *SIGMOD Record*, 33(4), 2004.
18. L. Van Elst, B. Klein, H. Maus, H. Schoning, A. Tommasi, C. Zavattari, J. Favaro, and V. Giannella. Business register interoperability throughout europe: The brite project. In *In Proc. The Semantic Web meets eGovernment, AAAI 2006*, Stanford University, California, March 2006.
19. T. Vitvar, A. Mocan, and V. Peristeras. Pan-european e-government services on the semantic web services. In *in Proc. WWW2006*, May 2006.
20. D. Weitzner, J. Hendler, T. Berners-Lee, and D. Connolly. *Creating a policy-aware Web: discretionary, rule-based access for the World Wide Web*. E. Ferrari and B. Thuraisingham (eds), Web and Information Security, Hershey, PA: Idea Group Inc, 2005.