Global Integration of Public Sector Information

Christos L. Koumenides, Manuel Salvadores, Harith Alani, and Nigel R. Shadbolt

School of Electronics and Computer Science,
University of Southampton,
SO17 1BJ, UK

{clk1v07, ms8, ha, nrs}@ecs.soton.ac.uk

ABSTRACT

This paper deals with technological methods for consolidating assets lists of available public sector information (PSI) for re-use. In this direction, the effort is to review the state of the art in delivering access to PSI throughout the world and to prioritize the necessary engagements for joining available PSI catalogues. We propose an architectural framework grounded on Semantic Web technologies to deliver a global platform for federated searching. A speculative survey of available PSI portals is presented, and the initial implementation, results, and analysis of the proposed architecture are covered in detail.

Keywords

Public Sector Information, Catalogue Integration, Linked Data

1. INTRODUCTION

Access to public sector information (PSI) plays a pivotal role in the world's information industry. The potential economic and social gain has created huge demand and pressure on governments [1-3]. At the same time, the evolution towards an open and unrestricted knowledge society influences the life of every citizen, by cultivating their rights to democratic participation and empowering them with new ways of accessing and acquiring knowledge. Recent surveys have shown that a vast majority of citizens feel that open access to information is important in "exercising their rights as citizens" [4]. Freedom of information distributes power to society by enabling people to make decisions about their public services (i.e. how their local schools and health services are performing), to have a say over the works of their representatives, and equally important, to have the right to innovate for better delivery and access to public information resources.

There are fundamental differences in the rules and methods each country employs to approach the subject of releasing public information for reuse. The UK and US have pioneered in this field by regulating strategies via their individual government agencies. Other governments are also following suit. At EU-level progress is monitored and supported by the transposition of key legislative frameworks; the 2003 PSI Re-use Directive [5], and the 2007 INSPIRE Directive [6].

Web Science Conf. 2010, April 26-27, 2010, Raleigh, NC, USA.

The absence of a single regulator and the need for clear consensus across countries, coupled with the need for cultural changes, will continue to hinder the development of an acceptable framework (legal or otherwise) of interoperability of a uniform PSI locator system. Despite this, a minimum stage of harmonization is desired if we are to respond to the growing demands for cross-border exploitation of information.

This paper deals with technological methods for consolidating assets lists of available public sector information for re-use. In this direction, the effort is to review the state of the art in delivering access to PSI throughout the world and to prioritize the necessary engagements for joining available PSI catalogues. We propose an architectural framework grounded on Semantic Web technologies to deliver a universal platform for federated searching.

The paper is structured as follows. Section 2 outlines our motivation for a global PSI locator system. Section 3 reviews initiatives around the world for delivering access to PSI. Section 4 proposes our architecture. Section 5 presents the implemented modules of the architecture. Section 6 deals with preliminary results and analysis. Section 7 contains a discussion on some of the more general issues related to current PSI locator services. Section 8 contains the conclusion.

2. MOTIVATION

Our proposal for a global PSI locator system is governed by four factual evidences:

1. Several countries throughout the world are beginning to publish their PSI catalogues on the Web
2. Studies indicate that PSI re-users engage in cross-border trading activities [2]
3. Absence of a centralized or otherwise globally coordinated PSI locator system
4. Several communities are republishing PSI as Linked Data (see Section 3.1)

The PSI Re-use Directive 2003/98/EC [5], which has now been broadly transposed by almost all EU Member States, clearly states the need for public sector information assets lists, preferably to be made accessible online via government portals. In addition, the European Commission's PSI Group is taking steps to monitor each Member State's progress with respect to delivering PSI Portals in a timely manner [7]. Article 9 of the Directive states:

"Member States shall ensure that practical arrangements are in place that facilitate the search for documents available for reuse, such as assets lists, accessible preferably online, of main documents, and portal sites that are linked to decentralized assets lists."

Additionally, recent reviews of the Directive have addressed the need for a universal metadata vocabulary. This has been the pinpoint of discussions during ePSIplus' thematic meetings in 2008 and 2009, in Latvia and Spain respectively [8, 9]. Section 3 reviews several examples of PSI locator services from around the world.

Furthermore, innovative uses of public sector information, such as navigation services, weather forecasts, real-time traffic control, industry and demographic services, often concern more than a single country or state. The need for public information resources, therefore, does not stop at the border. According to the MEPSIR study [2], indicators show that on average more than *8%* of total PSI re-uses involve some form of cross-border trading. The highest scores were indicative of countries with over *40%* of cross-border trading activities. Despite all evidence, there is currently no single comprehensive facility for retrieving cross-national datasets. Examples of early efforts are reviewed in Section 3.4, but these are by no means directed towards congregating all available worldwide facilities.

In support of evidence 4, we find several groups around the world eager to explore the use of Linked Data [10] to interlink and republish available PSI. Leading practitioners are indicated in Section 3.1. These efforts open a plethora of possibilities for global integration, as Linked Data allows the representation of data using URIs, which can easily be linked together. The paradigms of Linked Data PSI strongly motivate our proposition, since we may also be looking at a high level integration of Linked Data PSI from the catalogue/metadata perspective.

3. STATE OF THE ART

3.1 PSI Locator Services in US & UK

Locator services to public sector information are not a new concept. The United States have adopted an implementation of the GILS [11] framework in the mid-1990s to provide a single point of access to all US government information. The US Federal GILS employs standard network technologies and a set of approximately *70* metadata elements (GILS Core Elements) to describe and provide access to government resources in a uniform manner. The metadata terms cover information such as the description of the resource, information about its distributors, related dates, access constraints, identifiers, location of the asset, etc.

The Office of Public Sector Information (OPSI) in the UK has taken similar efforts in the late 1990s with the development of UK's Inforoute [12] service. Their efforts involved the development of a structured metadata catalogue for the key information resources that a public sector organization holds. These catalogues, also known as Information Asset Registers (IAR), are primarily records of unpublished government resources and cover virtually all types of information resources held by a public body. These include databases, old sets of files, recent electronic files, collections of statistics, research etc. The metadata elements used for creating IARs are drawn from the UK Government's Metadata Standard (eGMS) [13], which also includes entries pertaining to the Integrated Public Sector Vocabulary (IPSV) [14] standard (e.g. the "category" element). Furthermore, the UK government has been exploring the use of RDFa to enable greater re-use of IARs and interoperability with systems that may choose to harvest these records [15].

More recent and notable initiatives include the launching of the *data.gov* website in the US and the equivalent *data.gov.uk* in the UK. Both these technologies underpin the aforementioned IAR and GILS deployments and aim to produce records of large volumes of published PSI resources. The actual resources are hosted and provided either through the same websites or via other government portals, in which case a direct link to the recorded asset is provided. In addition, the *data.gov.uk* project aims to explore the potentials of the Semantic Web [16] for making data available in more transparent and open formats. After the appointment of Sir Tim Berners-Lee and Prof Nigel Shadbolt [17], the UK government has started to expose Linked Data end-points of available PSI for further re-use [18, 19]. Similar efforts are also seen in the works of the Renselear Polytechnic Institute (RPI) in the US, which has taken initiative to translate PSI datasets to RDF and generate RDFa catalogues through their Semantic Media Wiki extensions [20]. These have been major developments for the knowledge economy, empowering its citizens with an ever increasingly reusable form of its resources. Linked Data enables consumers to assimilate data from multiple sources in transparent and seamless ways. It is very likely that these technologies will have a drastic role in the evolution of the next generation of knowledge systems.

Similar initiatives in the UK are also surfacing from specific regions of the country, i.e. the *Pic and Mix* website[21], which aims to make Kent County Council's public data freely accessible in various open formats. The situation is very similar in the US, whereby several States have employed their own portals to PSI. Examples include New York City's *DataMine* website [22], the raw State data sources of the State of California, State of Utah, District of Columbia, and the City and County of San Francisco [23-26].

3.2 PSI Locator Services in Other EU Member States

Already several countries have started to promote a culture of re-use in the public sector. Information markets in Europe are beginning to shift their perspectives to more open access models, as opposed to the prevailing *cost-recovery* measures of the past [27]. During the 2008 review of the Directive [28] Member States had been assessed on their progress towards implementation. Arrangements that facilitate availability and re-use are beginning to surface throughout the European Union. Further reviews of the Directive will ensure that progress is indeed stable. This is strongly evident in efforts such as the Aporta [29] project in Spain, which aims to raise awareness among stakeholders and encourage discussion and debate around the subject. The project is foreseen to evolve into an adequate PSI portal for all public information needs of potential consumers. France, Belgium, and Italy have also indicated that progress is being made to expose their public data via proper national portals in the near future [30]. Sweden [31] and Finland [32] have opened countrywide portals with sufficient catalogues of available datasets.

3.3 PSI Locator Services in Other Regions and Continents

Several other countries, besides the EU and US, are producing and publishing their own PSI catalogues. Examples include the well-established portals of Australia [33], Russia [34], New Zealand's portal [35] to datasets available through its Government

Departments and Local Bodies, and Canada's various regional portals i.e. Vancouver's, Toronto's, and Nanaimo's [36-38].

3.4 European-level and Worldwide PSI Locator Services

Efforts to accumulate worldwide PSI resources are beginning to surface, as seen, for example, in the works of the Open Knowledge Foundation, CKAN [39]. CKAN is a registry of open data and aims to bring together packages that are open for public use. It currently hosts *768* packages indexed using a separate common cataloguing schema built on a few metadata terms.

Perhaps the most notable example of a European registry service is the recent opening of Eurostat's website [40]. Eurostat has been a pioneering service of European statistics since its creation in 1953, and became the European Commission's *Directorate-General* when the European Community was founded in 1958. Clearly, its aim is to provide an accurate picture of contemporary society in Europe, by bringing together and congesting statistics about the Member States. These include over *4000* statistical datasets, covering geographic, demographic, societal, economic, agricultural, and other forms. Available datasets are retrievable from the same website and can be browsed by theme, or alphabetically. Furthermore, datasets can be downloaded in bulk via Eurostat's bulk download facility. These are available in various formats, i.e. spreadsheets, tsv, xml, and html, and are typically accompanied by descriptive sets of metadata, which as a whole comprise an inclusive source of a European-level PSI catalogue.

Another example of a comprehensive source of global statistics is the World Bank developer's website [41]. The World Bank website offers a programmatic search facility across several indicators of key data sources from the World Bank association's accumulated statistics. However, the sources are searchable through the site's API and not via means of a standardized metadata catalogue.

The need for a coordinated global locator service is particularly obvious when resources from more than a single country or continent are concerned. For example, there is currently no single facility for retrieving the records of ethnographic or geographic resources residing at the discrete portals of different nations, i.e. the UK's and US'. One will need to contact several portals to locate all candidate resources. The following section presents our initial thoughts on an architectural framework to fulfill this gap. Our aim is to prioritize the necessary engagements for consolidating assets lists of available PSI.

4. PROPOSED ARCHITECTURE

Our proposed architecture leading to the integration of PSI catalogues is essentially a conceptual structure of interlaced activities and components. This is depicted in Figure 1.

In the first phase (Data Extraction and Linked Data layers) we download and transform a selection of catalogues with retrievable records to a common schema language format. We proceed with republication of the catalogues in a standard interface (see Section 5) and explore their degree of heterogeneity. Results will indicate the existence, or lack thereof, of common practices and unified standards exploited by the respective PSI portals.

In the second phase (Ontology Matching layer) we explore possible alignments between the catalogue schemata in the form of semantic correspondences between their entities. Schema matching [42] is a proposed solution to overcoming semantic heterogeneity problems. Many systematic approaches have been proposed in the field, and the literature spans across thousands of technical solutions. Although these share similar techniques (i.e. machine learning, statistical analysis) and exploit similar concepts (structure of ontologies, instance data, semantics), their degree of diversity is normally in the way they exploit and combine their results. Others differ in the degree of automation for producing their correspondences. A thorough investigation of available solutions will be necessary in order to develop a sound theory and implementation to fulfill our task.

Finally, the correct unification and interconnectivity of the catalogues will lead to federated searching and retrieval of records (Search Engine layer), which is our optimal goal.

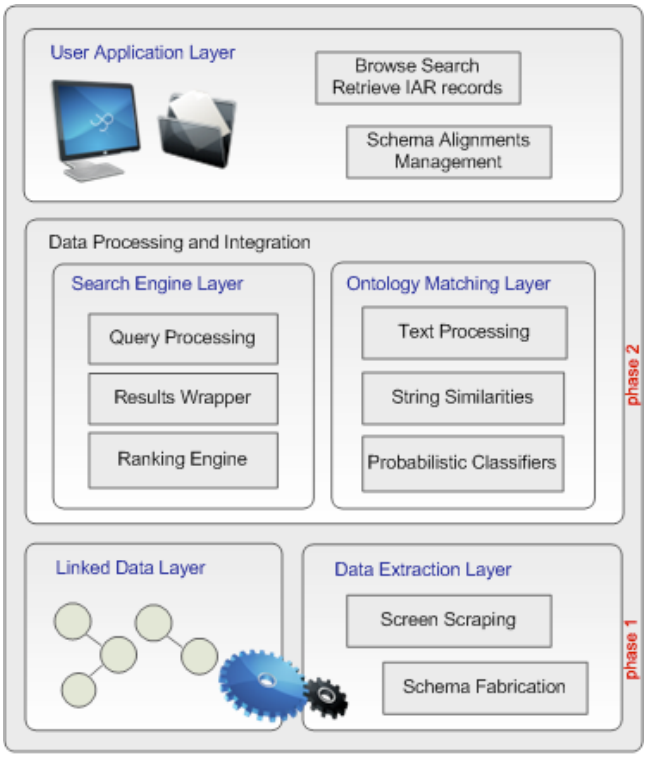

Figure 1: Activity design for unifying PSI catalogues

5. BUILDING A USE-CASE

This section reviews some of the completed modules of the aforementioned architecture. We focus on the modeling and re-fabrication of the PSI catalogues.

5.1 Selection of Catalogues

We identified four catalogues to start with the initial implementation of our integration scenario. These include the US RAW[1] and TOOL[2] sub-catalogues of *data.gov*, UK's national PSI catalogue at *data.gov.uk*[3], OPSI's Information Asset Registers (IAR)[4], and Australia's national catalogue at *australia.gov.au*[5].

[1] http://data.gov/catalog/raw (Accessed: 15 March 2010)

[2] http://data.gov/catalog/tools (Accessed: 15 March 2010)

[3] http://data.gov.uk/data (Accessed: 15 March 2010)

[4] http://opsi.gov.uk/iar (Accessed: 15 March 2010)

The selection of the catalogues has been a decisive choice, since it is necessary to consider catalogues of well-established portals, with sufficient amount of metadata. This gives us a variety of challenges to explore, ultimately leading to a more robust and generic integration engine. Table 1 lists the four catalogues, their access formats, and a brief quantitative overview of their contents.

Table 1: Overview of the selected PSI catalogues

	Format	Records	Metadata Elements
Data.gov (RAW & TOOL)	HTML/ CSV	1,563	20 approx.
Data.gov.uk	HTML/ CSV	3,002	25 approx.
Australia.gov.au	HTML/ RDFa	69	14 approx.
OPSI	HTML/ RDFa	2,514	16 approx.

5.2 Sourcing the Catalogues

The catalogues were sourced using the CSV files provided by the portals or via the development of custom web crawlers, which repeatedly requested records from the respective portals. Overall, we managed to download and extract the entire sets of records, except in very few cases where records contained malformed or misbehaving HTML code.

5.3 Modeling and Design Decisions

We decided to use RDF as the normal form for converting the catalogues to a common representation format. RDF enables the description of catalogue concepts and instance data as dereference-able URIs, which empower the datasets with a more reusable and highly potent format for cross-linking. RDF also ensures that the potential success of our framework will allow us to link our results back into the Linked Data cloud.

Envisioning the amount of data that our system will host we chose a scalable clustered triple store [43] and the ARQ extension from the Jena framework [44] to consume the data. ARQ provides a SPARQL programmatic interface for the Java programming language.

The initial translation has been intentionally kept minimal, just enough to reflect the original contents of the catalogues. In some cases, we employed universal namespaces to describe concepts/elements, although this has only been done in cases where the correspondence was particularly obvious (i.e. where we found a catalogue element named "description" we used the Dublin Core [45] "description" element in the conversion). Concepts, such as "agency", "publisher", and "department" appearing in the records were converted to first-class objects after a series of elementary text operations (case and blank normalization, punctuation elimination). Additionally, we made a choice to use the Tag Ontology [46] for representing all keyword concepts. The purpose of fine-tuning the catalogues this way is to enable the prospective for deeper linkage with external sources.

[5] http://data.australia.gov.au (Accessed: 15 March 2010)

5.4 Republishing the Catalogues

In order to align our work with the existing Web of Data and become part of the global data space, we chose to republish the catalogues on a standard Web interface. In doing so, we begin to embrace the four principles of publishing Linked Data on the Web, as set forth by Tim Berners-Lee [10] in 2006:

1. *Use URIs as names for things*
2. *Use HTTP URIs so that people can look up those names*
3. *When someone looks up a URI, provide useful information*
4. *Include links to other URIs, so that they can discover more things*

Intuitively, we developed a simple Web application using Java Server Pages for dereferencing catalogue URIs. The application is available online [6] and facilitates HTML and RDF content negotiation, according to the HTTP 303 status code. The resolution mechanism is based on a concise boundary description, whereby a request for an arbitrary URI will result in the execution and rendering of the following SPARQL query:

```
SELECT * WHERE {
?URI ?predicate-outbound ?object .
?subject ?predicate-inbound ?URI }
```

We extend the query accordingly to provide more data in the case of HTML requests (i.e. display available labels for each resource).

6. PRELIMINARY RESULTS & ANALYSIS

This section presents our initial analysis of the catalogues. In this preliminary study we are mainly looking at:

1. The temporal coverage of the catalogues
2. The main agencies in each country
3. The topic themes of the data

Our study begins by touching on some of the preliminary work necessary for making fundamental assertions about the contents. We start by focusing on the temporal alignment of the catalogues. The second part focuses on extrapolating the topic distribution.

6.1 Data Normalization & Time Series Analysis

The initial translation of the catalogues yields to a common high-level representation format, although the underlying instance data is more or less preserved in its original arrangement. There are several issues that demand to be resolved before we begin to analyze, compare, and eventually assert statements about the catalogues. One such issue is data normalization, which aims to extend the rudimental representation of the data to one that can be effectively compared and correlated. In this direction, data needs to be reconciled in order to adhere to a common representation standard.

As we begin to examine the catalogues, we realize that much of the temporal instance data (i.e. release dates, modification dates) is in fact not adhering to any universal standard. A universal format, however, is highly desired if we want to perform any form of time series analysis of the catalogues.

Let us consider an example of a record from the OPSI catalogue, with publication date "20090709". At first glance, the date

[6] http://bagatelles.ecs.soton.ac.uk/psi (Accessed: 15 March 2010)

appears relatively straightforward and can easily be translated to the ISO8601 standard as "2009-07-09", which matches the pattern *YYYY-MM-DD*. However, as we analyze the records of the respective catalogue we find several dates with values "20091501", or "200901". With well over *6,000* records of similarly ambiguous dates, a manual translation becomes highly inefficient. Our solution involves the development of automatic classifiers that extract the date parts (year, month, day) using a set of predefined rules[7]. The rules reflect common patterns found in the raw sources and attempt to match dates using those prototypes. We employ several rules for each model to accommodate for several patterns that indicate ambiguity.

The normalization process pilots a much cleaner output than what it was originally available. Using standardized dates we may begin to evaluate and compare several aspects of the catalogues. Figure 3 illustrates an example visualization of the size and temporal coverage of the catalogues according to their records' release dates[8]. We observe that OPSI is the most diachronic of the four, with records of resources published in as early as *1554*. *Data.gov.uk* (orange line) appears to be leading the way in both size and frequency throughout the years 2008-2010.

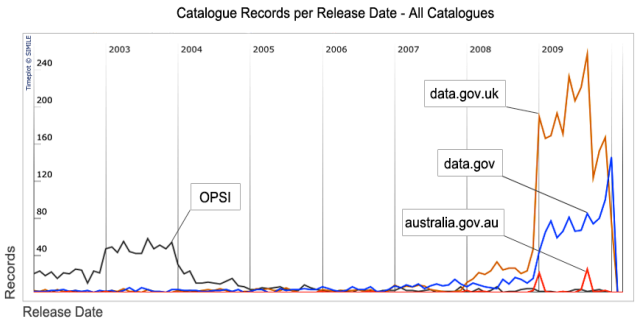

Figure 2: Catalogue records per release date – All catalogues

Additionally, we may plot the records of individual agencies against the main graphs. This gives us the opportunity to visualize the supply frequency and volume of assets disseminated by central government agencies. Figure 4 illustrates an example using the OPSI catalogue and one of its main distributors, the Driving Standards Agency. Table 3 lists the top distributors of recently published records for all the catalogues.

[7] The initial implementation has been completed, but further work will involve refinement of the algorithms.

[8] An online version to explore and refine the results is available: http://bagatelles.ecs.soton.ac.uk/psi/analyzer/release-dates/

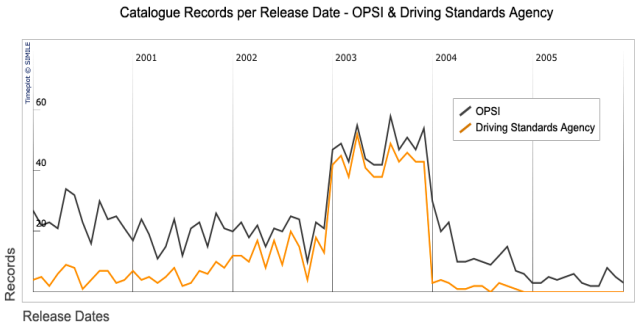

Figure 3: Catalogue records per release date – OPSI & Driving Standards Agency

Table 2: Top distributors in each catalogue

	Agency	Records
Data.gov (RAW & TOOL)	Environmental Protection Agency	462
Data.gov.uk	NHS Information Centre for health and social care	262
Australia.gov.au	Sustainability Victoria	5
OPSI	Driving Standards Agency	1,039

6.2 Data Topic Distribution

Further than the temporal coverage we are also interested in the topic themes of the data. There are several channels to guide this analysis. Some of the catalogues contain "category" concepts, which may tell us something about the general classification of the records. Likewise, we may focus on the "keyword", "title", or "description" concepts appearing in all of the catalogues. We are inclined to use a path that is available globally, and one that does not demand extensive filtering to furnish the general topic distribution. We settle on keywords, which are globally available and provide a high level perspective on the contents.

We are able to observe noteworthy variations in the four catalogues. Figures 4-5 present snapshots of the most frequently used keywords in *data.gov* and *data.gov.uk*. The contrast in topic distribution is remarkably evident. While in *data.gov.uk* we find terms such as "health" and "social care" to be the prevailing annotators, the focus switches to more environmental-related annotations when examining the *data.gov* catalogue, i.e. "toxic release", "chemical release", "facilities", etc. Australia's catalogue gives emphasis to "education", "environmental management", and generally themes related to environment and society. OPSI on the other hand concentrates on governmental affairs, i.e. "office services", "supplier contracts", "complaints", etc. The reader is directed to the online versions to explore the results[9]. Although keywords may not be adequate to indicate the incentives of publishers, they do offer a generalized picture of focal points of interest.

[9] http://bagatelles.ecs.soton.ac.uk/psi/analyzer/tags

Figure 4: Tag cloud of data.gov.uk

Figure 5: Tag cloud of data.gov

7. DISCUSSION

This section presents the outcomes of our preliminary research on the PSI portals outlined in Section 3. The results indicate several shortcomings with the existing platforms.

7.1 Accessibility and Standards

Plenty of the PSI channels we came across seem to have evolved their catalogues to serve internal needs, i.e. populate their own portals. Very few of the systems provided direct access to the raw formats of the catalogues (with the exception of *data.gov*, *data.gov.uk*, and *Eurostat*). On the contrary, catalogue entries are mostly discoverable via traditional HTML search facilities. This poses a significant challenge towards the development of a global integration facility, since most of the catalogues have to be crawled via programmatic means.

Additionally, if we are to truly maximise opportunities for exchange and re-use, providers must ensure that individual metadata standards move forward to the benefit of the society as a whole and not the direct communities that the information assets cover. While progress to facilitate access to PSI is monitored by initiatives and legislative frameworks at national and international levels, the extent of success in ensuring harmonisation of standards is still a factor dependent on local practices. The absence of clarity and harmonisation on rules and practises will without doubt continue to hinder the smooth functioning of a wider PSI society, in Europe and abroad. In this direction, the European Commission has been pressing on the issue of establishing a single metadata standard for all information assets of the Member States [8, 9]. Progress, however, is not yet evident. Very few countries with available PSI portals have yet to adopt a structured metadata standard, and one that is based on universal vocabularies. Moreover, many portals contain catalogue entries that are not inclusive and mainly consist of very few descriptive elements. Some countries do not even provide access points to their catalogues in a universal language, i.e. English, as the case is with Sweden and Russia. In other cases, we find a strongly structural and rich metadata at nationwide portals, but very poor and inconsistent metadata at the regional levels (as the case is with the US and UK portals).

7.2 Lack of Nationwide Access Facilities

While there is evidence of substantial progress throughout the world for opening PSI for re-use, many countries seem to lack on the organisational potency and discipline to provide their information assets via single nationwide portals. Access more often appears to be facilitated only by the local administrations that disseminate the information resources. This poses a significant barrier to awareness and cross-border discovery of PSI, consequently hindering much of its potential exploitation. We will surely expect more progress to come, as it is already obvious from several European and other studies. However, as a result of the current deficiency, our list of leading developments is limited to only the selected few indicated in Section 3.

8. CONCLUSION AND FUTURE WORK

The need to link PSI catalogues and data assets across jurisdictions emerges as a next step significant factor in the course of better and deeper exploitation of this national wealth. As governments continue to unlock the potentials of public sector information, either through their own initiatives or in response to market pressures, one should clearly expect more to come, not only in the volume and formats of the data, but also in the intelligence of its locator systems. It is of great importance to investigate the most appropriate platforms for digesting the sheer volume of cross-national datasets to aid potential re-users with their search requests.

This paper presented an architectural overview of the main processes for facilitating a global PSI locator system. A speculative survey of available PSI portals has been presented, and the initial implementation, results, and analysis of the proposed architecture were covered in detail. Our main line of future research involves the alignment of the PSI catalogues via semi-automatic means, to deliver a generic schema integration engine, and exploiting the alignments to facilitate federated searching.

9. AKNOWLEDGEMENTS

This work was supported by the EnAKTing project funded by the Engineering and Physical Sciences Research Council under contract EP/G008493/1.

10. REFERENCES

[1] G. Aichholzer, "Electronic access to public sector information: some key issues," in *Electronic Government. Third International Conference, EGOV 2004. Proceedings*, Zaragoza, Spain, 2004, pp. 525-528.

[2] M. Dekkers, *et al.*, "MEPSIR Study – Measuring European Public Sector Information Resources," ed: European Commission, 2006.

[3] E. Mayo and T. Steinberg, "The Power of Information," C. Office, Ed., ed, 2007.

[4] R. Marcella and G. Baxter, "Information need, information seeking behaviour and participation, with special reference to needs related to citizenship: results of a national survey," *Journal of Documentation,* vol. 56, pp. 136-160, 2002.

[5] C. European Parliament, "Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of public sector information," Directive COD 2002/0123 2003.

[6] C. European Parliament, "Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)," Directive COD 2004/0175 2007.

[7] E. Commission. (2009, 10 Jan 2010). *The importance of PSI Portals*. Available: http://www.epsiplatform.eu/news/news/the_importance_of_psi_portals

[8] B. Green, "PSI Asset Registers: towards a pan-European PSI registry," European Commission, Riga, Latvia 2008.

[9] B. Green, "PSI Asset Registers: towards a pan-European PSI registry," European Commission, Madrid, Spain 2009.

[10] C. Bizer*, et al.*, "Linked Data - The Story So Far," *International Journal on Semantic Web and Information Systems,* 2009.

[11] W. E. Moen, "The metadata approach to accessing government information," *Government Information Quarterly,* vol. 18, pp. 155-165, 2001.

[12] OPSI. *IAR and Inforoute*. Available: http://www.opsi.gov.uk/iar/index

[13] C. Office. *e-Government Metadata Standard*. Available: http://www.esd.org.uk/standards/egms/

[14] C. Office. *Integrated Public Sector Vocbulary*. Available: http://www.esd.org.uk/standards/ipsv/

[15] J. Sheridan, "The role of Information Asset Registers," ed: Office of Public Sector Information (OPSI), 2008.

[16] N. Shadbolt*, et al.*, "The Semantic Web revisited," *Ieee Intelligent Systems,* vol. 21, pp. 96-101, May-Jun 2006.

[17] T. Berners-Lee and N. R. Shadbolt, "Put in your postcode, out comes the data," *The Times,* 2009.

[18] T. Omitola*, et al.*, "Put in your postcode, out comes the data: A Case Study," presented at the 7th Extended Semantic Web Conference, Greece, 2010.

[19] H. Alani*, et al.*, "Unlocking the Potential of Public Sector Information with Semantic Web Technology," presented at the The 6th International Semantic Web Conference (ISWC), Busan, Korea, 2007.

[20] L. Ding*, et al.*, "The Data-gov Wiki: A Semantic Web Portal for Linked Government Data," in *ISWC*, 2009.

[21] KCC. *Pic and Mix, a KCC Innovation Project*. Available: http://picandmix.org.uk/categories/

[22] NYC. *NYC DataMine*. Available: http://nyc.gov/html/datamine/html/data/data.shtml

[23] *Data - State of California*. Available: http://www.ca.gov/data

[24] *Data - Utah.gov*. Available: http://data.utah.gov/

[25] *Data Catalog - District of Columbia*. Available: http://data.octo.dc.gov/

[26] *DataSF - Liberating City Data*. Available: http://datasf.org/

[27] P. Weiss, "Borders in Cyberspace: Conflicting Public Sector Information Policies and their Economic Impacts," ed: U.S. Department of Commerce, 2004.

[28] E. Commission, "Re-use of Public Sector Information - Review of Directive 2003/98/EC," ed, 2008.

[29] *Proyecto Aporta*. Available: http://www.proyectoaporta.es/web/guest/index

[30] E. Commission, "Minutes of the 1st Meeting on Public Sector Information Portals, Luxembourg, 25 September 2009," 2009.

[31] *Offentliga datakallor - opengov.se*. Available: http://www.opengov.se/sidor/english/

[32] *Laatua Verkkoon*. Available: http://suomi.fi/suomifi/laatuaverkkoon/suomifi_verkosto/kansalaisosallistujan_tyokalut_-kilpailu_2009/

[33] *Catalogue - data.australia.gov.au*. Available: http://data.australia.gov.au/catalogue

[34] *OpenGovData*. Available: http://opengovdata.ru/

[35] *Open Data Catalogue - Open Data, Open Government*. Available: http://cat.open.org.nz/category/dataset/

[36] *City of Vancouver Open Data Catalogue*. Available: http://data.vancouver.ca/datacatalogue

[37] *toronto.ca - Open - Building a city that thinks like the web*. Available: http://www.toronto.ca/open/

[38] *Nanaimo Data Catalog*. Available: http://www.nanaimo.ca/datafeeds/

[39] *CKAN - Comprehensive Knowledge Archive Network*. Available: http://www.ckan.net/

[40] *Eurostat - Your Key to European Statistics*. Available: http://epp.eurostat.ec.europa.eu/portal/page/portal/eurostat/home/

[41] *The World Bank Developer Network*. Available: http://developer.worldbank.org/page

[42] J. Euzenat and P. Shvaiko, *Ontology Matching*: Springer-Verlag New York, Inc., 2007.

[43] *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).*

[44] *Jena Semantic Web Framework*. Available: http://jena.sourceforge.net/

[45] *DCMI - Dublin Core Metadata Initiative*. Available: http://dublincore.org/

[46] *Tag Ontology*. Available: http://www.holygoat.co.uk/projects/tags/