A classification scheme for open government data: towards linking decentralised data

Evangelos Kalampokis*

University of Macedonia,
156, Egnatia str., 54006, Thessaloniki, Greece
Fax: +30-2310-891-509
E-mail: ekal@uom.gr
*Corresponding author

Efthimios Tambouris

Department of Technology and Management,
University of Macedonia,
Periohi Loggou-Tourpali, 59200, Naousa, Greece
Fax: +30-2332-052-462
E-mail: tambouris@uom.gr

Konstantinos Tarabanis

Department of Business Administration,
University of Macedonia,
156, Egnatia str., 54006, Thessaloniki, Greece
Fax: +30-2310-891-544
E-mail: kat@uom.gr

Abstract: Open government data (OGD) refers to making public sector information freely available in open formats and ways that enable public access and facilitate exploitation. Lately, a large number of OGD initiatives launched worldwide aiming to implement one-stop portals acting as single points of access to governmental data. At the same time, the so-called linked data technologies emerged aiming at publishing structured data on the web in such a way that enables semantically enriching data, uniform access to data, and linking of data. In this paper, we first propose a classification scheme for OGD initiatives based on the relevant literature. We thereafter, review and analyse OGD initiatives based on the proposed scheme. We finally present an architecture and prototype implementation for the most advanced OGD class in our scheme, which enables linking decentralised data.

Keywords: open government data; OGD; classification; linked data; web of data; architecture; one-stop government data portals.

Reference to this paper should be made as follows: Kalampokis, E., Tambouris, E. and Tarabanis, K. (xxxx) 'A classification scheme for open government data: towards linking decentralised data', *Int. J. Web Engineering and Technology*, Vol. X, No. Y, pp.000–000.

Biographical notes: Evangelos Kalampokis is a PhD student at the University of Macedonia, Greece and an Associate Researcher at the Information Systems Laboratory at the same university. He is also a member of the Linked Data Research Centre at DERI Galway, Ireland. He holds a Diploma in Electrical and Computer Engineering in 2003 from the Aristotle University of Thessaloniki, Greece and an MSc in Business Administration in 2006 from the University of Macedonia, Greece. His main research interests include, but are not limited to linked data and semantic web technologies as well as the domains of e-government and e-participation.

Efthimios Tambouris is an Assistant Professor at the Department of Technology and Management, University of Macedonia, Thessaloniki, Greece. Prior to that, he served at research centres CERTH/ITI and NCSR 'Demokritos' and the IT industry. He holds a Diploma in Electrical Engineering from the National Technical University of Athens, Greece and an MSc and PhD from Brunel University, UK. During the last 12 years, he has initiated, managed and participated in numerous eGovernment research projects, service contracts and standardisation activities. He has more than 90 peer-reviewed publications in e-government and e-participation.

Konstantinos Tarabanis is a Professor at the Department of Business Administration of the University of Macedonia, Greece; and the Director of the Information Systems Laboratory at the same university. He received an Engineering Diploma in Mechanical Engineering from the National Technical University of Athens; an MS in both Mechanical Engineering and Computer Science; and a PhD in Computer Science from Columbia University, New York, NY. He was a research staff member at the IBM T.J. Watson Research Centre (1991–1994). He has more than 150 peer-reviewed publications in software modelling and development for the domains of e-government, e-business, e-learning, e-manufacturing, etc.

1 Introduction

Public sector collects, produces and disseminates a wealth of information ranging from financial statistics to every day incident reports. This information can be an important primary material for added value services and products, which can increase government transparency, improve public administration's function, contribute to economic growth and provide social value to citizens (Burdon, 2009; Newberry et al., 2008; Dekkers et al., 2006). The importance of government data reuse and exploitation is evident when considering that relevant policies already exist for some years, e.g., Directive 2003/98/EC of the European Commission (European Commission, 2003) and Freedom of Information and Paperwork Reproduction Acts in the USA (Gellman, 2004). This commitment has been recently re-enforced, e.g., the government of the USA committed that it "*will take appropriate action, consistent with law and policy, to disclose information rapidly in forms that the public can readily find and use*" (Obama, 2009). The government of the UK committed to "*release valuable public datasets and make them free for reuse*" (HM Government, 2009). In addition, the European Union considered the "*increase of the availability of public sector information for reuse*" as one of its most important objective as regards electronic government (e-government) (Ministers responsible for e-government policy of the European Union Member State, 2009).

In the last couple of years, a large number of governments worldwide started to massively make data available on the web. This *open government data* (OGD) movement follows the open data philosophy (Ayers, 2007) suggesting making data freely available to everyone, without limiting restrictions. It is based on the publication of data in open formats and ways that make it accessible and readily available to the public and allow reuse (Alonso et al., 2009). One of the main tenets of OGD is that government provides data and then private parties built added value products and services that provide interactive access for the public (Robinson et al., 2009). From a technological point of view, linked data technology has recently emerged to facilitate amongst others uniform access and linking of data that reside in different datasets.

The main objectives of this paper are three. Firstly, to suggest a classification scheme based on relevant literature. Secondly, to identify, analyse and classify OGD initiatives using the proposed scheme. We believe that having a classification scheme allows a deeper understanding of initiatives and therefore the domain as a whole. Finally, to propose an architecture and prototype implementation for the most technologically advanced class in the scheme, which suggests linking decentralised data.

The remaining of this paper is organised as follows. In Section 2, we review related work. In Section 3, we propose the classification scheme and analyse its relevant classes. In Section 4, we analyse and classify real-life OGD initiatives worldwide. In Section 5, we present an architecture and prototype implementation for the most advanced class in our scheme. Finally, in Section 6 conclusions are drawn along with future work.

2 Related work

As already suggested, OGD initiatives emerged only recently. In many cases however OGD is part of e-government and more specifically online one-stop government and governmental portals. Therefore, in this section we first present work related to e-government emphasising online one-stop government data portals and relevant classification schemes as then proceed with work related to OGD. Finally, we present challenges of OGD activities due to data management.

2.1 E-government emphasising one-stop government

During the last decade, a number of models and schemes have been suggested by international organisations, consulting firms and researchers for describing the development of e-government. Layne and Lee (2001) introduced a ‘stage of growth’ model comprising four stages, namely cataloguing, transaction, vertical integration and horizontal integration. These stages are explained in terms of organisational and technological complexity as well as different levels of integration. Based on this model, Andersen and Henriksen (2006) proposed the public sector process rebuilding model. Here, the key dimensions are the degree of activity-centric websites and processing of the end-users information and service requests. In addition, Siau and Long (2005) developed a five-stage model, which is described in terms of time, complexity and integration as well as benefits and costs. More specifically, according to this model time spending, system complexity, integration, benefits and costs all increase with the advancement of e-government. Finally, Lee (2010) suggested a common frame of reference for

e-government development models. This framework comprises two dimensions, namely citizen/service perspective and operation/technology perspective.

A significant part of e-government initiatives involves the development of governmental portals acting as single point of access to governments. One-stop government portals face amongst others legal, organisational, technological and cultural challenges (Tambouris, 2001; Dias and Rafael, 2007). In this paper, we concentrate on organisational aspects arising from the establishment of a portal acting as a single point of access as opposed to making data available from each public agency's website.

For our analysis, we employ as a starting point the data manufacturing systems paradigm where the production and storage of data has been conceptualised (Strong et al., 1997). In data manufacturing systems, three roles are identified and each role is associated with a process: data producers are associated with data production process; data custodians with data storage, maintenance and security; and data consumers with data utilisation processes, which may involve additional data aggregation and integration.

In the case of OGD initiatives the three roles become:

R1 customer (which is actually the data consumer)

R2 one-stop government data portal (portal)

R3 public agency (which is actually the data producer).

Furthermore, the main data-related processes are:

P1 data production process

P2 data utilisation processes

P3 data publishing and maintenance process

P4 data aggregation and integration process

P5 data searching process

P6 data collection process.

2.2 OGD and linked data

OGD initiatives emerged only recently and as a result, there is a lack of academic studies to analyse them based on classification schemes. There is however an increasing number of practical guidelines suggested by various stakeholders. We present in this sub-section two sets of guidelines which in our view constitute the most influential approaches.

The World Wide Web Consortium (W3C) e-Government Interest Group suggest three steps for public administrations to open and share their data (W3C, 2009):

- firstly, publish data in raw form by means of files in well-known and non-proprietary formats such as CSV and XML
- next, create online catalogues of the raw data
- finally, make the data machine-readable.

Sir Tim Berners-Lee invited governments to publish data according to linked data principles (Berners-Lee, 2009). He further proposed a five-star maturity model as follows (Berners-Lee, 2010):

1 star: publishing data on the web even in proprietary and desktop-centric formats

3 stars: publishing data in machine-readable formats such as spreadsheets documents

3 stars: publishing data in machine readable and non-proprietary formats using open standards, e.g., CSV

4 stars: publishing data using linked data principles

5 stars: linking the available data.

It should be noted that the UK Government committed to publish most of its non-personal data as linked data by the end of 2011 (HM Government, 2009).

As 'linked data' seem to have a prevalent role in the future of OGD initiatives we briefly introduce the relevant technologies. The term linked data refers to "*data published on the web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets*" (Bizer et al., 2009). Linked data is based on semantic web philosophy and technologies but in contrast to the full-fledged semantic web vision, it is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inferencing (Hausenblas, 2009). It promises the creation of the 'web of data' as data from decentralised and heterogeneous sources can be interlinked through typed links. Web of data aims at replacing isolated data islands with a giant distributed dataset built on top of the web architecture (Heath, 2008).

Linked Data following a RESTful approach requires the identification of resources with URI references that can be dereferenced over the HTTP protocol into RDF data that describes the identified resource. In addition, Linked Data include the creation of typed links between URI references, so that one can discover more data. More specifically, the four linked data principles as described by Sir Tim Berners-Lee (2010) are the following:

- all item should be identified using URIs
- all URIs should be dereferenceable, that is, using HTTP URIs allows looking up the item identified through the URI
- when looking up a URI it leads to more data, which is usually referred to as the follow your nose principle
- links to other URIs should be included in order to enable the discovery of more data.

Linked data distinguishes between information and non-information resources (Bizer et al., 2008). The former refers to all the resources we find on the traditional document web such as documents, images etc, while the latter refers to real world thing such as people, schools, laws, public agencies etc. The adoption of identifiers ensures to uniquely identify information resources in the web but not the real world things the information resources refer to Halpin (2006). Hence a central issue in the web of data is the finding of identifiers that refer to the same real world thing[1] (Jaffri et al., 2008). These identifiers became known as *URI aliases* (Bizer et al., 2008).

The use of linked data technologies for publishing data on the web provides the following advantages:

- Enables data to be integrated with the web[2]. This describes the ability to link together different pieces of information published on the web and the ability to directly reference a specific piece of information.
- Reduces the challenge of integrating heterogeneous data and building large-scale, ad hoc mashups (Heath, 2008; Hausenblas, 2009).

Based on the main activities required for publishing linked data (Bizer et al., 2008), we derive three more data-related processes, which are present in OGD initiatives adopting a linked data approach:

P1 URI definition and management process

P2 data vocabulary creation and management

P3 data links creation and management.

2.3 Challenges of OGD initiatives due to data management

The review of the literature presented in the previous sub-sections suggests OGD activities involve three main roles and nine processes relevant to data and metadata management. It is important to note that unlike the data manufacturing systems paradigm where each process is assigned to one actor, in OGD initiatives more options exist and the decisions might have wider consequences. The main processes and relevant options include:

- Who own (i.e., maintains) the data? This could be either the public agency or the portal. This is particularly important as it relates directly to data quality, e.g., data may become obsolete is not properly and timely maintained.
- Who publishes the data (and possibly related metadata)? This could be either the public agency or the portal or both. Again, this decision might have consequences in data quality. In case both public agencies and the portal publish data, it is possible that different values appear in different places due to lack of proper synchronisation.

It should be also noted that the above-mentioned decisions can be driven by different factors. An important factor is efficiency and effectiveness. Other factors include the principles of subsidiarity (which suggests that matters ought to be handled by the smallest, lowest or least centralised competent authority), legitimacy, transparency, accountability and trust. These are particularly important considerations in the case of public sector and therefore should be also considered in the case of OGD initiatives.

Finally, we should note that other challenges also exist when designing OGD initiatives. For example, a main challenge associated with establishing a main portal is which agency will host this portal. Other considerations include the shift of power which might be present when there is a change in the authority that own public data. These are both strategic decisions but are outside the scope of this paper hence will not be further considered.

3 OGD classification scheme

The proposed classification scheme includes two dimensions. The first dimension cares for the technological aspect of OGD initiatives, which is an important driver. In Sub-section 2.2, we have already presented purely technology-driven schemes that can be used for classifying OGD initiatives (although it should be admitted that they were not initially proposed as classification schemes). The second dimension cares for the fact that OGD activities are actually online one-stop government data portals thus non-technical, domain-specific peculiarities should be also considered. In summary, the proposed classification scheme comprises the following dimensions:

- the technological approach followed for making data available on the web
- the organisational approach followed for data provision.

Each dimension is elaborated before presenting the proposed classification scheme.

The first dimension refers to the technological approach followed. In general there are many different technological approaches for making data available on the web such as

a as downloadable files in proprietary formats

b through custom APIs

c as downloadable files in machine readable formats

d through RESTful APIs

e through search interfaces.

These are characterised by how easy it is to

a use the relevant technology

b access the data over the web

c extract and reuse data

d link together different pieces of information.

The proposed scheme includes two broad technological approaches. These are further divided according to Tim Berners-Lee five-star technological maturity model described in the Sub-section 2.2.

- The first approach suggests making data available on the web as downloadable files in well-known formats such as PDF, Excel, CSV, KML, XML, JSON etc. This broad category is further divided based on the format in:
    1 making data available on the web as downloadable files in proprietary and desktop-centric formats, e.g., PDF
    2 making data available on the web as downloadable files in machine-readable formats, e.g., Excel
    3 making data available on the web as downloadable files in machine-readable formats using open standards, e.g., CSV.

- The second approach suggests making data available on the web as linked data through RESTful APIs and/or SPARQL search interfaces. This broad category is further divided in:
  1 making data available based on Linked Data principles (i.e., HTTP, URIs and RDF)
  2 linking data from different datasets.

The second dimension of the proposed classification scheme is related to the organisational approach followed for providing governmental data. In Sub-section 2.1, we reviewed several e-government models that can be used as classification schemes while in Sub-section 2.3, we highlighted the challenges of OGD initiatives. Based on this analysis, we conclude that an adequate classification of ODG initiatives should include the organisational approach they are following for providing data. Broadly speaking, two approaches are possible:

- Data belonging to various public agencies is published by the one-stop government data portal. We use the term *direct data provision* to dictate this method of data provision.
- Data belonging to various public agencies is published in a decentralised manner by these agencies (usually in their website) while the portal provides some kind of linking mechanism and/or metadata for the identification of the actual dataset. We use the term *indirect data provision* to dictate this method of data provision.

Clearly, other approaches can be also considered for covering the organisational aspects of OGD initiatives. We believe however that selecting the data provision approach enables us to address most of the public sector considerations presented in Sub-section 2.3 in a simple and straightforward manner.

Figure 1 The proposed OGD classification scheme (see online version for colours)

<table>
<tr><th colspan="2">Technological Approach / Organizational Approach</th><th>Direct Data Provision</th><th>Indirect Data Provision</th></tr>
<tr><td rowspan="3">Downloadable Files</td><td>Proprietary and desktop-centric formats</td><td rowspan="3">Repository of Downloadable Files</td><td rowspan="3">Registry of Downloadable Files</td></tr>
<tr><td>Machine-readable formats</td></tr>
<tr><td>Machine-readable formats using open standards</td></tr>
<tr><td rowspan="2">Linked Data</td><td>Linked data principles</td><td rowspan="2">Direct Provision of Linked Data</td><td rowspan="2">Indirect Provision of Linked Data</td></tr>
<tr><td>Linking available data</td></tr>
</table>

By combining the technological and organisational dimensions we derive the classification scheme shown in Figure 1.

We now outline the main characteristics of each of the four main dimensions.

3.1 Downloadable files

The main advantage is that data is provided in simple to use formats that are widely accepted by both developers and customers, e.g., citizens and businesses.

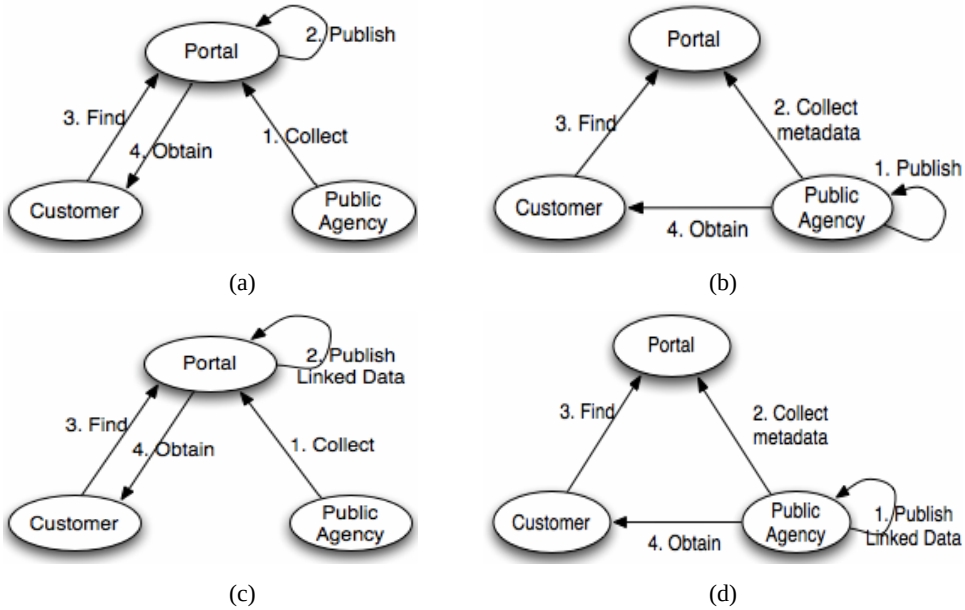
3.2 Linked data

The use of linked data technologies infuses the technical advantages of linked data, i.e., ability to link to a specific piece of data and reusing part of the data. On the other hand, the effort and time needed for publishing the data (i.e., finding vocabularies, assign URIs etc.) are large while at the same time the relevant technologies are still immature and not widely adopted. Furthermore, technological challenges still exist such as those related to standardised querying of distributed data sources using SPARQL (Hartig et al., 2009).

3.3 Direct data provision

Having all data at one place (portal), suggests that aggregated, processed and value-added data can be provided by the governmental portal. On the other hand, maintainability is limited, e.g., in cases where data change over time, an efficient and effective data synchronisation process must be in place to prevent the portal from providing obsolete data.

Figure 2 Main data processes in the classes of the proposed scheme, (a) repository of downloadable files (b) registry of down downloadable files (c) direct provision of linked data (d) indirect provision of linked data



3.4 Indirect data provision

The fact that the actual data is published by the data producer itself means that the provided data is the only one of its kind (unique) and also up to date. These characteristics contribute to the increase of data believability and data accuracy. On the other hand, aggregated, processed and value data cannot be provided by the portal; if this is needed, it has to be performed by the customer.

In Section 2, we identified three main roles (customer, portal, public agency) and nine different data-related processes. Figure 2 shows the main data-related processes in each class of the proposed scheme, namely repository of downloadable files, registry of downloadable files, direct provision of linked data, and indirect provision of linked data. Figure 2 clearly illustrates that the same data process can be performed by different actors as suggested in Sub-section 2.3.

Table 1 List of the identified OGD initiatives

Name	*URL*	*Responsible authority*
Catálogo de Datos de Asturias	http://risp.asturias.es	The Principality of Asturias, Spain
City of Edmonton Open Data catalogue	http://data.edmonton.ca/	City of Edmonton, Canada
Data.australia.gov.au	http://data.australia.gov.au	The Australian Government
Data.ca.gov	http://data.ca.gov	State of California
Data.gov	http://data.gov	The Federal US Government
Data.gov.uk	http://data.gov.uk	The UK Government
Data.govt.nz	http://data.govt.nz	New Zealand government
Data.nsw	http://data.nsw.gov.au	New South Wales Government
Data.seattle.gov	http://data.seattle.gov	The Seattle City Government
Data.vic.gov.au	http://data.vic.gov.au	The Victorian Government
DataSF	http://www.datasf.org	The City of San Francisco
Dati.piemonte.it	http://www.dati.piemonte.it	Region of Piedmont, Italy
Dc.gov data catalogue	http://data.octo.dc.gov	District of Columbia
Lichfield Open Data	http://lichfielddc.gov.uk/data	Lichfield District, UK
London datastore	http://data.london.gov.uk	Greater London Authority
Maine.gov DataShare	http://www.maine.gov/data	State of Maine
Mass.gov/data	http://mass.gov/data	The Commonwealth of Massachusetts
NYC Data Mine	http://nyc.gov/html/datamine	The City of New York
OpendataNI	http://www.opendatani.info	Northern Ireland, UK
Open Data Euskadi	http://opendata.euskadi.net	Basque Country, Spain
Pic and Mix	http://picandmix.org.uk/	Kent County, UK
Toronto.ca/open	http://toronto.ca/open	The City of Toronto, Canada
Vancouver's Open Data Catalogue	http://data.vancouver.ca/	The City of Vancouver, Canada
Warwickshire Open Data	http://opendata.warwickshire.gov.uk/	County of Warwickshire, UK

4 Review of OGD initiatives

In this section official OGD initiatives are reviewed. In Table 1, the names and the URLs of the initiatives are presented along with the responsible authorities. In total 24 OGD initiatives were identified.

We note that this review includes initiatives commenced by public administrations and those aiming to create a portal acting as a single point of access for data originated and produced by disparate pubic agencies

The analysis of the initiatives revealed some interesting results. In terms of governance level, analysis suggests initiatives have been established in federal (e.g., the Federal US government and the Australian government), national (e.g., the UK government), regional (e.g., the State of California and the Victorian government) and local (e.g., the City of Vancouver and the District of Columbia) level.

In terms of objectives, analysis suggests data included in the identified initiatives contribute towards most of the declared objectives of OGD, i.e., enhance transparency, enable economic growth, improve citizens' every day life and support public administration's function. More specifically, transparency can be enhanced by datasets including, e.g., governmental spends, financial statements and statistics and building permits. This type of data is provided by some of the initiatives such as data.gov.uk and data.govt.nz. Economic growth can be achieved by the liberation of datasets including geo-spatial data and/or census statistics data. Social value to citizens can be provided by datasets describing the location of schools, bus stops, hospitals etc., crime incidents by location, available social workers and meals programs for homeless. This sort of data is the most common one and appears in the majority of the initiatives. Finally, the function of public administration can be supported by datasets related to legislation and the organisational structure of public sector.

In terms of technological advancement, analysis suggests three of the identified initiatives publish their data according to linked data principles, namely data.gov in the USA, data.gov.uk in the UK and the data catalogue of the Principality of Asturias in Spain. Data.gov.uk seems to be the most advanced initiative and the one concentrating the most interesting characteristics.

In Figure 3, the identified OGD initiatives are classified according to the proposed scheme. The majority of the initiatives fall in the first class (i.e., direct data provision based on downloadable files) while the third class (i.e., centralised provision of linked data) includes only three initiatives. In addition, there is only one initiative (i.e., data.gov.uk) falling in the fourth class characterised by the indirect provision of linked data. We should also note that a number of initiatives fall into more than one categories. This is due to the fact that these initiatives use more than one technological approach. For example they provide data in both proprietary and open formats. However, we should underline that the organisational approach followed is always the same with data.gov.uk being the only exception.

Data.gov.uk includes the indirect provision of data in machine-readable formats using both proprietary and open formats as well as the provision of linked data in both a direct and indirect manner. In our understanding, the provision of linked data by different public agencies such as the HM Treasury, the Department of Education and the Office for National Statistics using sub-domains of data.gov.uk (e.g., education.data.gov.uk) denotes direct data provision. But we consider the provision of geo-spatial data by

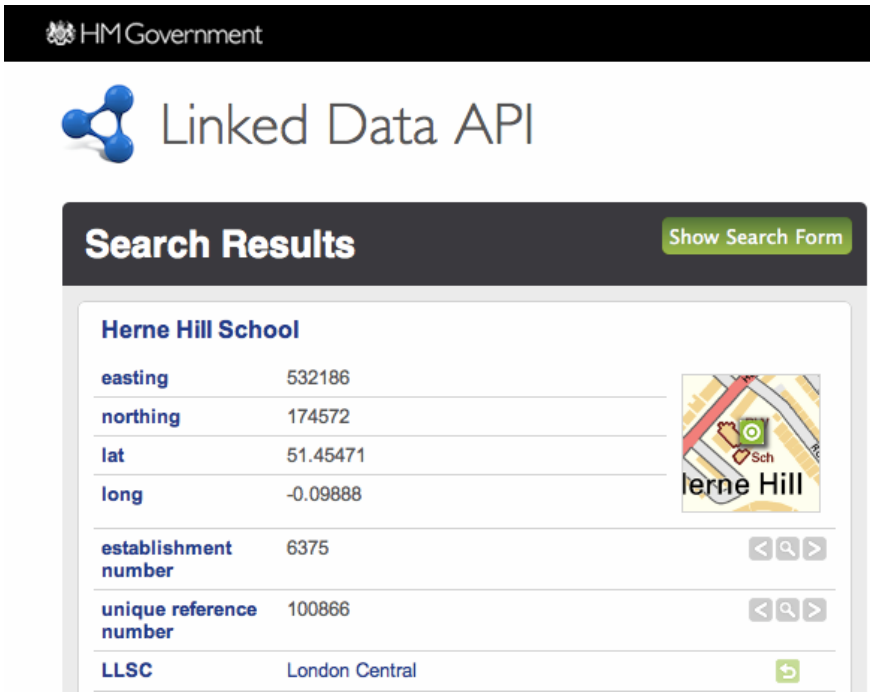
ordnance survey as *indirect provision of linked data* because it is performed in the agency's official URL.

Data.gov.uk is the only initiative included in the *Linking available data* sub-class as links between different datasets have been created. In Figure 4, an example of this linking is depicted. More specifically, data from the Department of Education describing schools is linked to data from the Office for National Statistics. The 'joint point' of these datasets is the Local Learning Skills Council (LLSC) that is responsible for the specific school.

Figure 3 OGD initiatives grouped according to the classification scheme (see online version for colours)

Technological Approach \ Organizational Approach		Direct Data Provision	Indirect Data Provision
Downloadable Files	Proprietary and desktop-centric formats		Data.govt.nz Data.nsw Mass.gov/data Open Data Euskadi
	Machine-readable formats	NUC data mine Pic and Mix Toronto.ca/opne Vancouver's open data catalogu	Data.australia.gov.au Data.ca.gov Data.gov.uk Data.govt.nz Data.nsw Data.vic.gov.au DataSF OpendataNI
	Machine-readable formats using open standards	City of Edmonton open data catalogue Data.gov Data.seattle.gov Dati.piemonte.it Dc.gov - data catalogue Lichfield open data London datastore Maine.gov DataShare Toronto.ca/open Vancouver's open data catalogue Warwickshire open data	Data.ca.gov Data.gov.uk Data.vic.gov.au
Linked Data	Linked data principles	Catalogo de Datos de Asturias Data.gov Data.gov.uk	
	Linking available data		Data.gov.uk

Figure 4 Screen-shot of data.gov.uk site showing the linking of data from the Department of Education (schools) to data from the Statistics Office (admin areas) (see online version for colours)



5 Indirect provision of linked data: architecture and prototype implementation

The analysis presented earlier suggests only one OGD initiative (namely data.gov.uk) adopts linked data and a partially indirect data provision approach. We have already shown that an indirect data provision approach features some interesting public sector characteristics (trust, accountability etc.) and that Linked Data as a technology can fully support it but poses additional considerations (see Section 2). However, we could not find an OGD initiative that fully adopts this operation model.

In this section, we present architecture and a prototype implementation for indirect provision of Linked Data. This enables investigating whether this approach is technically feasible and what are the technical considerations. For doing so, we take a number of decisions (e.g., who manages URIs etc.). We acknowledge other decisions are possible.

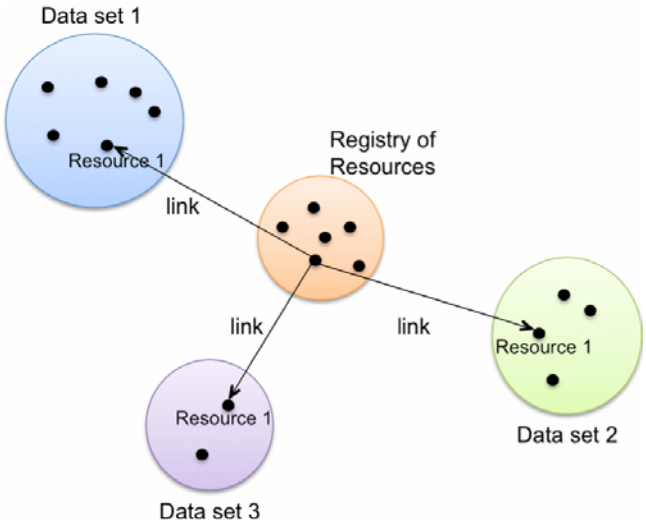
The proposed architecture is particularly useful for linking data that reside in different datasets and actually belong to different public agencies that can even belong to different government sectors (e.g., health, education etc.). This is achieved through the creation of links between URI aliases in a specific sector. In this case, the portal is also responsible:

- to maintain a list of the available resources
- to publish this list as linked data and thus assign a URI to each resource

- to identify URI aliases between this list and each new published government dataset
- to create links between these URI aliases.

The portal can be therefore also considered as a Registry of Resources (RoR) since it maintains a list of available resources in the area and assigns a URI to each of them. Figure 5 depicts a data-level view of the architecture. This figure illustrates the role of the RoR as a single point of reference for all public agencies that provide information about a specific resource.

Figure 5 A data-level view of the architecture (see online version for colours)



We now present a prototype implementation of the proposed architecture. In this scenario we consider three actors:

- a school (namely Moraitis): this publishes information about itself on its website
- the 2nd Local Directorate of Secondary Education of Athens (local directorate onwards): this maintains a relational database with information about all schools in its area (including Moraitis) and publishes it on its website
- the Ministry of Education: this maintains a relational database with information about all schools in the country and publishes it on its website.

We assume the ministry acts as one-stop government data portal for education thus in our scenario it hosts the RoR.

The steps that are followed in this prototype implementation are five.

1 First, local directorate decides on a vocabulary for describing the schools it supervises. We assume this is based on well-known ontologies such as FOAF and SKOS and is denoted by the namespace ‘ath’. Although the implementation details are out of the scope of this paper, the main tasks are:
   - creating a RESTful API for the dataset (e.g., the URI of Moraitis school is http://195.251.218.37:2020/resource/publicentity/259)

- publishing a SPARQL endpoint for the dataset (http://195.251.218.37:2020/sparql).

2 Second, local directorate publishes its data (currently in a relational database) as Linked Data. For this purpose, we used D2R server (Bizer and Cyganiak, 2006) since it is one of the most mature relevant solutions[3].

3 Third, Moraitis School uploads a FOAF file on its web space for describing itself. It should be noted that although FOAF is mainly used to describe people, it can be also used to describe organisations (since both foaf:Person and foaf:Organization classes are sub-classes of foaf:Agent class).

4 Fourth, RoR publishes a list of all Greek schools as linked data. For doing so, we inserted school name and unique ID into a relational database and used D2R server to publish the information as linked data; this is denoted by the namespace 'ror'. The relevant tasks include:

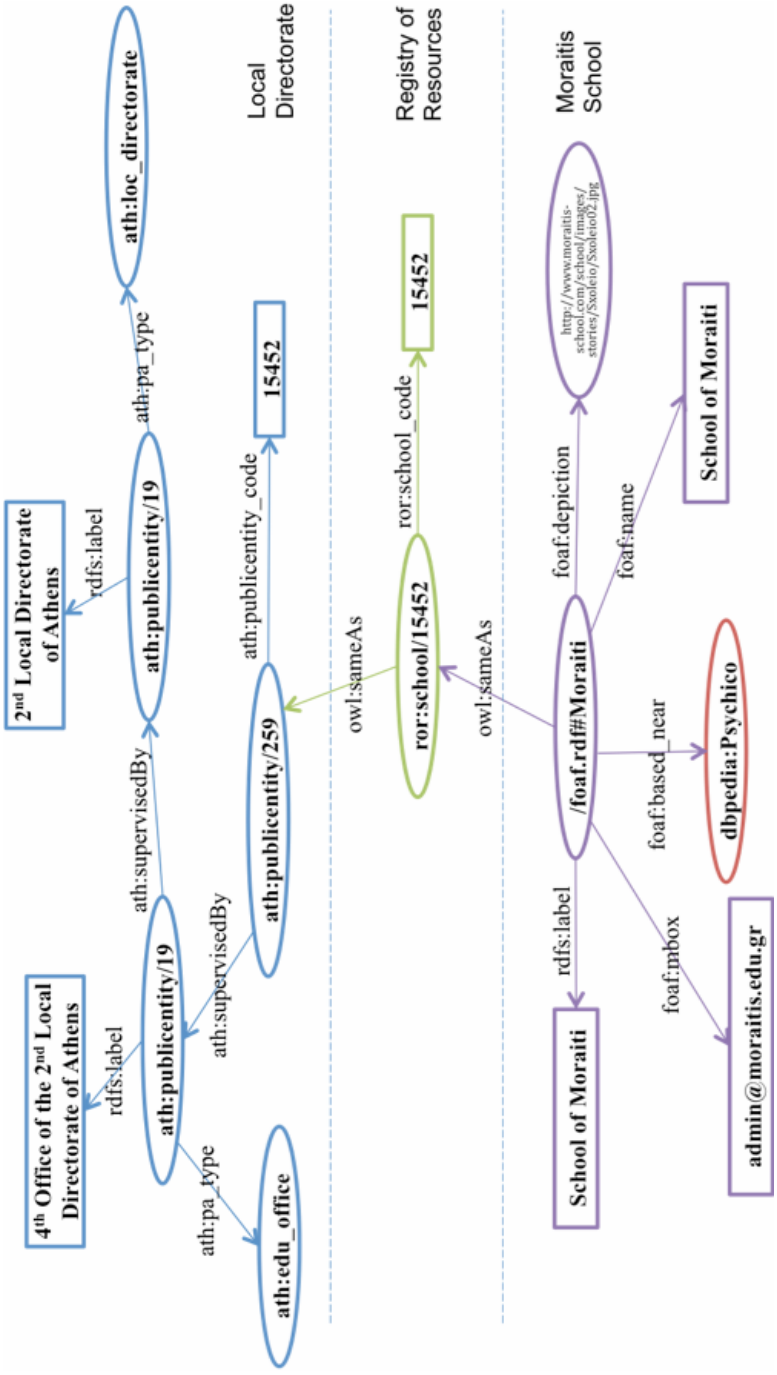
- assigning a dereferencable URI to each school (e.g., http://195.251.218.39:2020/resource/school/15452)
- creating a SPARQL endpoint for having access to the specific dataset (http://195.251.218.39:2020/sparql).

5 Fifth, using the owl:sameAs[4] predicate and the Silk framework (Volz et al., 2009), we identified URI aliases between the RoR and the two pubic agencies and we created links between them. Silk is accessing the datasets we want to link together through their SPARQL endpoints and calculates the similarity values between specific values of the resources described in the datasets. In our implementation we used the unique school ID aiming to link the datasets provided by the local directorate and by the RoR. The result is a CSV file containing the links above an accepted threshold (in our case 98%). Finally we imported the CSV file in the relational database of the RoR. As a result the dataset provided by the RoR and the dataset provided by the local directorate are connected through owl:sameAs links between URI aliases. We should note here that we assume the vocabulary used by the target dataset (i.e., the local directorate) was known to the source dataset (i.e., the RoR). This assumption enables the values comparison of *ath:publicentity_code* and *roe:school_code* properties.

The linking of the resource described by the FOAF file and the respective resource in the RoR was done manually during the creation of the FOAF file. More specifically, the FOAF file indicates that the resource that describes is *owl:sameAs* the entity 'Moraitis School' described in the RoR (ror:school/15452). In general, using this approach we were able to publish information on 859 schools as Linked Data.

The final outcome of these steps as regards Moraitis is the graph depicted in Figure 6. This Figure presents the linking of unique data about the specific school provided by two distributed sources. The *ror:school/15452* representation of the school is provided by the RoR and is the glue between the representations *ath:publicentity/259* and */foaf.rdf#Moraiti* since these two are linked to the first one by two *owl:sameAs* links. More specifically, the school publishes information such as name, e-mail address, location, a picture of the school etc while local directorate publishes information about the specific office of the directorate that is responsible for supervising Moraitis School.

Figure 6 The linked data graph (see online version for colours)



By following the specific approach, the data provided by different public sources about Moraitis School is now linked and the data customer can search for and get an integrated view by using semantic mashup tools such as Sig.ma (Tummarello et al., 2010). In addition, due to the typed links between the disparate sources of government data one can follow these links and receive more relevant information such as other schools in the area.

6 Discussion and conclusions

OGD refers to making public sector information freely available in open formats and ways that enable public access and facilitate exploitation. OGD is a political priority for many countries worldwide and is related to other similar priorities such as e-government and online one-stop governmental portals. Lately, OGD has been reinforced also due to technological advances in Linked Data, which facilitate publishing structured data on the web in such a way that enables semantically enriching data, uniform access to data, and linking of data.

In this paper, we reviewed the literature in OGD, linked data, e-government maturity models and online one-stop governmental portal. Based on this review, we propose a classification scheme with two main dimensions.

The first dimension refers to the technological approach followed for making data available on the web. This includes two main categories

a making data available of the web as downloadable files in well-known formats such as PDF, Excel, CSV, KML, XML, JSON etc.

b making data available of the web as Linked Data through RESTful APIs and/or SPARQL search interfaces.

The first approach is technologically mature but does not enable automatic access at the data level or linking of data.

The second dimension refers to the organisational approach followed for providing governmental data. Again, this includes two categories

a direct data provision, where data belonging to various public agencies is published by the one-stop government data portal

b indirect data provision, where data belonging to various public agencies is published in a decentralised manner by these agencies (usually in their website) while the portal provides some kind of linking mechanism and/or metadata for the identification of the actual dataset.

The first approach enables data aggregation at the portal while the second enables data ownership and management by public agencies, which is sometimes related to the principles of subsidiarity, legitimacy, transparency, trust and accountability. These principles are often important for the public sector and might even have priority over efficiency and effectiveness.

Based on the proposed classification scheme, we reviewed 24 official OGD initiatives. The analysis results suggest they span various administrative levels (from national to local), they support various political objectives (such as transparency,

accountability etc.) and the data are mainly provided as downloadable files (with the exception of three initiatives).

We further proposed a technical architecture for the forth class in the classification scheme, which includes indirect provision of linked data. The goal is to link disparate government data sources through the creation of *owl:sameAs* links between URI aliases in a specific sector (e.g., education). For doing so, the portal maintains a list of all available resources, assigns a dereferencable URI to each of them, and identifies URI aliases between this list and other government linked datasets. Finally we applied the proposed architecture to implement a prototype scenario about the consumption of data related to a specific school that is provided by two different public agencies. This exercise showed the approach is technically possible albeit immature (e.g., it requires significant manual handling).

Future work includes further investigation of the potential and limitations of indirect provision of linked data in the public sector from an organisational and technological perspective. For example, technological challenges include streamlining the operation of the RoR by automating the process of linking URI aliases between different sources as well as enabling complex and distributed queries over available linked data provided by disparate government sources. Organisational aspects under further investigation include better understanding the relationship between data provision models and political priorities, such as accountability, transparency, legitimacy, trust and the like.

Acknowledgements

The authors would like to express their gratitude to Michael Hausenblas for his valuable comments and support as well as the anonymous referees for valuable comments that allowed substantially improving the paper.

References

Alonso, J. et al. (2009) ‘Improving access to government through better use of the web’, available at http://www.w3.org/TR/2009/NOTE-egov-improving-20090512.

Andersen, K.V. and Henriksen, H.Z. (2006) ‘E-government maturity models: extension of the Layne and Lee model’, *Government Information Quarterly*, Vol. 23, pp.236–248.

Ayers, D. (2007) ‘Evolving the link’, *IEEE Internet Computing*, Vol. 11, No. 3, p.96, pp.94–95.

Berners-Lee, T. (2009) ‘Putting government data online’, available at http://www.w3.org/DesignIssues/GovData.html.

Berners-Lee, T. (2010) ‘Design issues: linked data’, available at http://www.w3.org/DesignIssues/LinkedData.html.

Bizer, C. and Cyganiak, R. (2006) ‘D2R server – publishing relational databases on the semantic web’, *Poster in the 5th International Semantic Web Conference (ISWC2006)*, Athens, Georgia, USA.

Bizer, C., Cyganiak, R. and Heath, T. (2008) ‘How to publish linked data on the web’, available at http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial.

Bizer, C., Heath, T. and Berners-Lee, T. (2009) ‘Linked data – the story so far’, *International Journal on Semantic Web and Information Systems (IJSWIS)*, Vol. 5, No. 3, pp.1–22, Special Issue on Linked Data.

Burdon, M. (2009) 'Commercializing public sector information privacy and security concerns', *Technology and Society Magazine*, Vol. 28, No. 1, pp.34–40, IEEE.

Dekkers, M., Polman, F., Velde, R. and de Vries, M. (2006) *MEPSIR: Measuring European Public Sector Information Resources*, Final Report of Study on Exploitation of public sector information – benchmarking of EU framework conditions, available at http://ec.europa.eu/information_society/policy/psi/docs/pdfs/mepsir/final_report.pdf.

Dias, G.P. and Rafael, J.A. (2007) 'A simple model and a distributed architecture for realizing one-stop e-government', *Electronic Commerce Research and Applications*, Vol. 6, No. 1, pp.81–90.

European Commission (2003) 'Directive 2003/98/EC of the European parliament and of the council on the re-use of public sector information', *Official Journal of the European Union*, Vol. 345, pp.90–96.

Gellman, R. (2004) 'The foundations of United States Government information dissemination policy', in Aichholzer, G. and Burkert, H. (Eds.): *Public Sector Information in the Digital Age: Between Markets, Public Management and Citizens' Rights*, pp.123–136, Ed-ward Elgar, Cheltenham, UK.

Halpin, H. (2006) 'Identity, reference, and meaning on the web', *Proceedings of the Workshop on Identity, Meaning and the Web (IMW06)*, Edinburg, UK.

Hartig, O., Bizer, C. and Freytag, J.C. (2009) 'Executing SPARQL queries over the web of linked data', in Bernstein, A. et al. (Eds.): *ISWC 2009*, pp.293–309, LNCS 5823.

Hausenblas, M. (2009) 'Exploiting linked data to build web applications', *IEEE Internet Computing*, Vol. 13, No. 4, pp.68–73.

Heath, T. (2008) 'How will we interact with the web of data?', *IEEE Internet Computing*, Vol. 12, No. 5, pp.88–91.

HM Government (2009) 'Putting the frontline first: smarter government', available at http://www.hmg.gov.uk/media/52788/smarter-government-final.pdf.

Jaffri, A., Glaser, H. and Millard, I.C. (2008) 'URI disambiguation in the context of linked data', *Proceedings of the 1st Workshop on Linked Data on the Web (LDOW2008)*, Beijing, China.

Layne, K. and Lee, J. (2001) 'Developing fully functional e-government: a four stage model', *Government Information Quarterly*, Vol. 18, No. 2, pp.122–136.

Lee, J. (2010) '10 year retrospect on stage models of e-government: a qualitative meta-synthesis', *Government Information Quarterly*, Vol. 27, No. 3, pp.220–230.

Newberry, D., Bently, L. and Pollock, R. (2008) 'models of public sector information provision via trading funds', available at http://www.berr.gov.uk/files/file45136.pdf.

Obama, B. (2009) 'Transparency and open government', *Memorandum to the Heads of Executive Departments and Agencies*, available at http://www.whitehouse.gov/the_press_office/Transparency_and_Open_Government.

Robinson, D., Yu, H., Zeller, W.P. and Feltern, E.W. (2009) 'Government data and the invisible hand', *Yale Journal of Law and Technology*, Vol. 11, p.160.

Siau, K. and Long, Y. (2005) 'Synthesizing e-government stagemodels: ameta-synthesis based on meta-ethnography approach', *Industrial Management & Data Systems*, Vol. 105, No. 4, pp.443–458.

Strong, D.M., Lee, Y.W. and Wang, R.Y. (1997) 'Data quality in context', *Communications of the ACM*, Vol. 40, No. 5, pp.103–110.

Tambouris, E. (2001) 'An integrated platform for realising online one-stop government: the eGOV project', *Proceedings of the 12th International Workshop on Database and Expert Systems Applications (DEXA 2001)*, Munich, Germany.

Tummarello, G., Cyganiak, R., Catasta, M., Danielczyk, S., Delbru, R. and Decker, S. (2010) 'Sig.ma: live views on the web of data', *Proceedings of the 19th International World Wide Web Conference (WWW2010)*, Raleigh, North Carolina, USA.

Volz, J., Bizer, C., Gaedke, M. and Kobilarov, G. (2009) 'Silk – a link discovery framework for the web of data', *Proceedings of the 2nd Linked Data on the Web Workshop (LDOW2009)*, Madrid, Spain.

W3C (2009) 'Publishing open government data', W3C Working draft, available at http://www.w3.org/TR/gov-data/.

W3C RDB2RDF Incubator Group (2009) 'Survey of current approaches for mapping of relational databases to RDF', available at http://www.w3.org/2005/Incubator/rdb2rdf/RDB2RDF_SurveyReport.pdf.

Notes

1 We should note that in the rest of this paper we refer to non-information resources by using the term resource.

2 http://webofdata.wordpress.com/2010/03/01/data-and-the-web-choices/

3 For a complete review of relevant tools and approach consider (W3C RDB2RDF Incubator Group, 2009)

4 http://www.w3.org/TR/owl-ref/#sameAs-def