**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

# A State-of-the-Art Analysis of the Current Public Data Landscape from a Functional, Semantic and Technical Perspective

**Michael Petychakis[1], Olga Vasileiou[1], Charilaos Georgis[1], Spiros Mouzakitis[1], John Psarras[1]**

[1] National Technical University of Athens, School of Electrical and Computer Engineering, Athens, Greece, mpetyx@epu.ntua.gr, olgaa148@gmail.com, c.georgis@lse.ac.uk, smouzakitis@epu.ntua.gr, john@epu.ntua.gr

Received 12 August; received in revised form 7 November 2013; accepted 18 December 2013

**Abstract**

Open Government Data initiatives and particularly Open Government Data portals have proliferated since the late 2000's. A comprehensive analysis of the capabilities and potential of these initiatives is currently missing from the recent research literature. In order to address this gap, the paper at hand aims towards analyzing the landscape of Open Governmental Data in the European Union from a functional, semantic and technical perspective. Our research focused on the collection and categorization of an indicative number of public data sources for each of the 27 European Union country-members through investigating their services and characteristics. By modeling and classifying the data sources according to their key attributes, we were able to proceed to their statistical analysis and assessment in terms of their content, licensing, multilingual support, acquisition, ease of access, provision and data format. Our results portray the current quality of Public Sector Information infrastructures and highlight what still needs to be done in order to make public data truly open and readily available for researchers, citizens, companies and innovation in general.

**Keywords:** Public sector information, National open data portals, Open governmental data, Semantic interoperability, Standardization

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile
This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

# 1 Introduction

Since the late nineties there has been an explosion of activity around open data and especially open government data [12]. Under the official recommendations and support of the European Commission (EC) [6], national governments all over the European Union (EU), invest on the establishment of national open data portals [2] in order to increase public access to high value, machine readable datasets generated by public agencies and organizations [13]. In 2003, the EU adopted the Directive on the re-use of public sector information (PSI Directive) that introduced a common legislative framework regulating how public sector bodies should make their information available for re-use in order to remove barriers such as discriminatory practices, monopoly markets and a lack of transparency [6]. All 27 EU Member States are directed to implement the PSI Directive via national legal orders. On April 10th of 2013, the EC announced that the European Union Member States have approved a text for the new PSI Directive. The new Directive highlights the importance of re-usability of public data and further encourages the proliferation of open data portals [22].

Since the official launch of national open data repositories, their impact has been thoroughly examined by relevant literature. Bertot et al. [1] identified transparency, trust and deeper citizen relationship with the public interest, as key long-term benefits of open data portals. On the other hand, Hogge [8] mentions the criticism that US and UK's (Site 1) data portals have attracted with regard to their actual impact on the daily lives of citizens. Moreover, Hogge [8] investigates data strategies and characteristics of open data portals and addresses non-tackled issues for developing and middle income countries. Additionally, Gurstein [7] takes a supportive but critical look at open government data from the perspective of its possible impact on the poor and marginalized and concludes that there may be cause for concern in the absence of specific measures being taken to ensure that there are supports for ensuring a wide basis of opportunity for *effective data use*. Meanwhile, Maier and Huber [14] discuss the impact of open data on the relations among citizens, public administration, and political authority, while Ding et al [5] highlight potential innovative aspects of open data portals and lessons learned. Throughout literature, interoperability amongst government data sources is a common issue addressed, especially at the semantic level. For instance, Janssen [12] et al. emphasize on the importance of interoperability in e-government infrastructures and data portals. The semantics and language that each data portal is tied to, is one of the most common and inherent interoperability challenges [9]. The majority of the datasets is published in the native language of the country that maintains the data portal. Another major interoperability challenge lies within the wide range of non-compatible licenses that each country promotes, as explained by Miller et al. [17].

Another interesting topic that is closely related to the open data landscape is open innovation, which is one of the hottest topics during the past decade. There is a variety of reasons behind that and there is also extensive literature describing in detail the context, the circumstances under a firm could use such innovation and in general the outcomes available [10], [16], [20].

Open data portals are a great resource for innovation, especially when dealing with linked open data; mainly because of their connected nature [24]. More specifically, when standards like Resource Description Framework (RDF) are applied, then through the SPARQL Protocol and RDF Query Language (SPARQL) there is a common Application Programming Interface (API) for all the applications who would like to facilitate this data. In general, open data is available for developers and researchers to build their ideas on top of those. So, the need for common practices and frameworks could rapidly increase the adoption of such technologies. Although it was a common attitude that it is expensive for organizations to adopt open data because of their cost, there are several counter examples on that thesis [14] and we can see more and more companies which actually facilitate such technologies.

Innovation through open standards can increasingly boost creativity in both enterprise communities as well as start-ups, which are already a well-established community worldwide. For example, there is a current trend where such initiatives are focused around location data. New technologies that utilize location such as smartphones, tablets, even watches and others are widely used and more and more software is being developed around those. To succeed in those directions developers are trying to reuse information available regarding such data. Big companies like Nokia, Google, and Facebook are providing their own data through APIs and developers utilizing them, build software applications.

Similar efforts are being made in the Linked Geodata (Site 5) and the Open GeoSpatial (Site 6) which also provides the (Site 7). Those efforts provide more or less the same information as the enterprise tools listed above in a common SPARQL interface, whereas the rest use their own APIs and their own protocols, and have mostly pay as you go policies. An analysis of this is given by Raivio and Luukkainen [19] while they explore the open innovation benefits in the mobile area.

Extending the above discussion, there are also open data portals complying with the same open standards like RDF and SPARQL, providing maps for areas uploaded by simple users and not just from institutions or companies. This use case is a success story in the State of Oregon regarding the marine board (Site 8). Providing incentives, such as game prizes, to citizens and young developers can have a significant impact on the crowd-based enrichment of the

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

open data repositories. During the last years, a vast number of open data communities that develop new ideas and apps, have emerged around Open Government data portals.

The proliferation of such Open Government Data initiatives and particularly Open Government Data portals during the recent years, however, has raised significant questions in terms of the variety in the interoperability, standards, services, openness and vision of each portal. A systematic analysis of these portals is only limitedly available [25]. A comprehensive analysis of the capabilities and potential of these initiatives is currently missing from the recent research literature.

In order to address the above gap this paper utilizes the characteristics and suppositions of the related literature in order to perform a state-of-the-art analysis of public data infrastructure from a functional, semantic and technical perspective so as to gain a better insight into the current public data landscape. More specifically, we focused on the countries of the European Union, where we intended to identify irregularities and challenges with regard to the provision of Open Data across the studied countries. The research began with a snapshot of the current situation of open data sources in Europe [4]. Thereafter, we managed to collect, categorize, statistically analyze and comparatively assess the open government data throughout the European Union.

# 2 Methodology

Within the current study, a representative number of diverse and distributed open government data sources from all the countries of the European Union were analyzed. Data was collected by means of an online research for organizations and public offices of each country that provide open public data in electronic form, as well as data aggregators. The collected datasets were classified and analyzed by genre and country in view of enabling a qualitative comparison between the EU member states [18].

Our research mainly focused on a central government level, meaning that for each of the 27 EU countries we reviewed every working ministry website with publicly available datasets. Our research was also extended to other central government-related websites, such as national and regional open data portals, public services, national statistical offices, central banks, national geodata-related websites as well as each country's official police, fire service and army website. Furthermore, besides national public datasets, we also included official European Union portals and websites.

In addition to this research, which targeted at covering an in-breadth analysis, we also conducted in-depth research for 3 representative countries of the European Union that have exhibited a proliferation of open data initiatives during the last years. For that purpose, we chose the United Kingdom (being a pioneer about open data amongst European countries), France (being a representative country of mainland Western Europe) and Greece (being a country of Eastern Europe).

With regard to this research:

- The United Kingdom demonstrates great differentiation among local administration structures. Unlike other European countries, there are a great number of local administrative sections, a tradition holding since the Middle-Ages [21]. In terms of this research, counties, boroughs and unitary authorities, were regarded as equivalent to the municipalities in the rest of Europe. In summary, we proceeded to the investigation of 98 counties, 61 boroughs, 15 unitary authorities and 60 urban areas (cities/towns).

- In France, we investigated 95 departments' and 11 urban areas' (cities/towns) data sources [3].

- In Greece, 13 Regions and 325 Municipalities were investigated, as predefined by the *Kallikrates* program of the Greek Ministry of Internal Affairs (Site 4).

We subsequently proceeded to the collection and categorization of the datasets from each data source researched.

The process was initiated on March 2012 and was concluded on January 2013. In total, 3,466 datasets were found.

The open government data sources were categorized and investigated based on the following attributes [15]:

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile
This paper is available online at www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

Table 1: Attributes of the collected datasets

| Attribute | Type | Description |
|---|---|---|
| Country | Semantic | The country that the dataset originated from. |
| Data Source Type | Semantic | The Data Sources have been categorized per type as follows: Ministry, National Open Data Portals, CKAN Initiative, Central Bank, Data Aggregators, Law, Regulatory Authority, National Statistical Office, Statistical Office, Public Service, City, Municipality, Borough, County, Community, Department, Region, Unitary Authority, District. |
| General Title | Semantic | A title for the data source. |
| Subtitle | Semantic | A subtitle for the data source. |
| Uniform Resource Locator (URL) | Semantic | The URL for the data source. |
| Author | Semantic | The initial author of the public dataset. |
| Author E-mail | Semantic | The author's email. |
| Maintainer / Publisher | Semantic | The organization responsible for publishing the dataset. |
| License | Functional / Semantic | The License of the dataset (e.g. Open Government License UK, Creative Commons). |
| Category | Semantic | One or more category themes that the dataset belongs to. The categorization according to the content of the datasets was: [Arts and Recreation], [Business Enterprise, Economics, and Trade], [Budget, Revenues & Expenditures], [Construction, Housing, and Public Works], [Crime and Community Safety], [Demographics], [Education], [Elections], [Emergency Services], [Energy and Utilities], [Environment, Geography and Meteorological], [Health and Disability], [Labor Force and Employment Market], [Law Enforcement, Courts, and Prisons], [Political], [Tourism], [Urban Transport], [Defense], [Multi / Various]. |
| Type of Information | Semantic | A brief description of the data source category. |
| Period | Semantic | The chronological period concerning the dataset. |
| State of Data | Technical | Whether the information provided is static or dynamic (e.g. real-time data). |
| Coverage | Semantic | The regional span covered by each dataset. |
| Catalog/ Discover | Functional / Technical | The different ways to browse for specific information throughout the website; these are referred to as follows: Free text search, Browse of categories, Comprehensive Knowledge Archive Network (CKAN, Field-based search, Filters, Map / Spatial, SPARQL Search. |
| Data Acquisition | Functional / Technical | How the provider has generated / produced the data available in each data source: Internal - Survey / Research, Internal - Back office / Everyday service operations, External – Harvesting, External - Uploaded by Public Agencies, External - Uploaded by Users. |
| Data Provision | Functional / Technical | The set of services offered to the user (Online View of Dataset / Download file/ Charts/ Map/ API). |
| Feedback | Functional / Technical | Feedback mechanisms: Request Dataset forms, Rate Datasets, View popular demands / vote best data requests, and Comment on Datasets. |
| Language Interface | Semantic | The language(s) the user interface is available in. |
| Language Data | Semantic | The language(s) the datasets themselves are available in. |
| Data Format | Technical | The format of the available datasets (Excel/ PDF/ CSV just to name a few). |
| Metadata | Semantic | If available, the metadata [23] standard for the data catalog of the data source. |

In summary, 1629 data sources were investigated and a more detailed overview of those in relation with the collected datasets is as follows: 1580 datasets came from Ministries, 904 from Municipalities, and 573 from Regions. As far as Public Services were concerned, 171 datasets were collected. 65 came from Law and Regulatory Authorities, 56 from Data Aggregators, 45 for Central Banks and 34 from National Statistical Offices. Finally, 14

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile
This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

datasets were gathered from Open Governmental Portal Websites, 15 from Ckan Initiative Websites and 8 datasets from various Statistical Offices.

# 3 Results

In this section, the results of the analysis per key attributes are presented.

## 3.1 Licenses

One of the most important issues of public data is the definition of the license related to it and whether it is open or not. The following chart (Figure 1) portrays in absolute numbers the distribution of the various licenses encountered during the collection of the datasets:

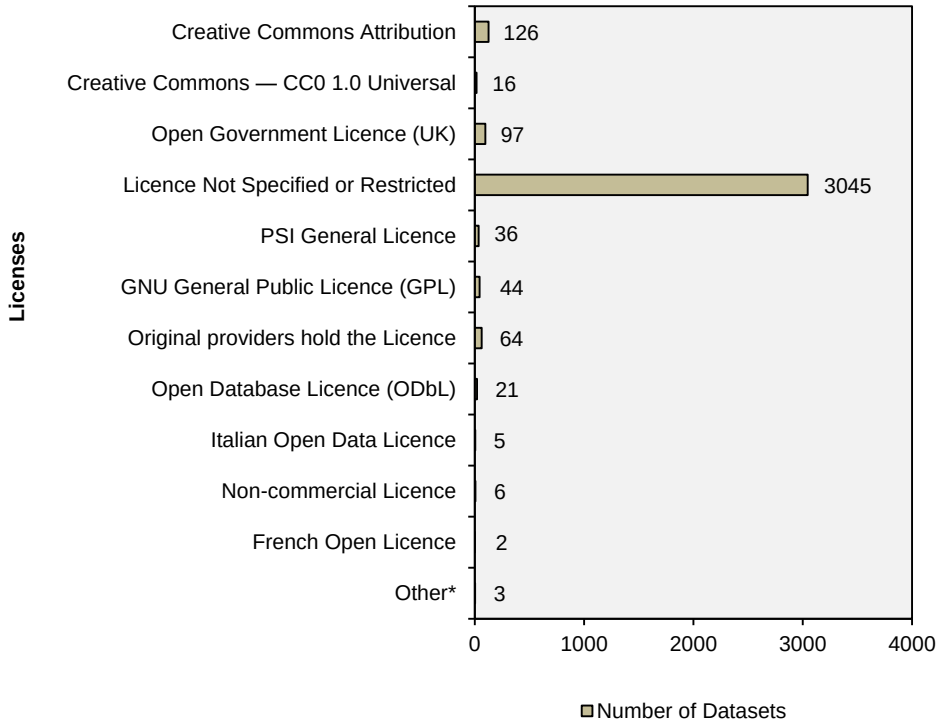



Figure 1: Licenses of the analyzed data sources

The chart shows that the vast majority of the datasets (3,045 or 87.88% of all of the datasets) are published without a clearly defined or open license (License Not Specified or Restricted) while in the case of 64 of them (1.85%) the original providers held the license. On the contrary, only 356 datasets (10.27%) were published with an open license, with most of them under either the UK Open Government License or the Creative Commons Attribution License.

It is evident that one of the most crucial steps into making data truly open – apply an open license that will support the openness of data- is now lacking in Europe, with the bright exception of the emerging national open data portals. In addition, in order to discern which countries have made steps forward towards opening their data and which of them are most lacking in that aspect we compared the data from the 27 European countries that were examined and reached the following conclusions for each of them:

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

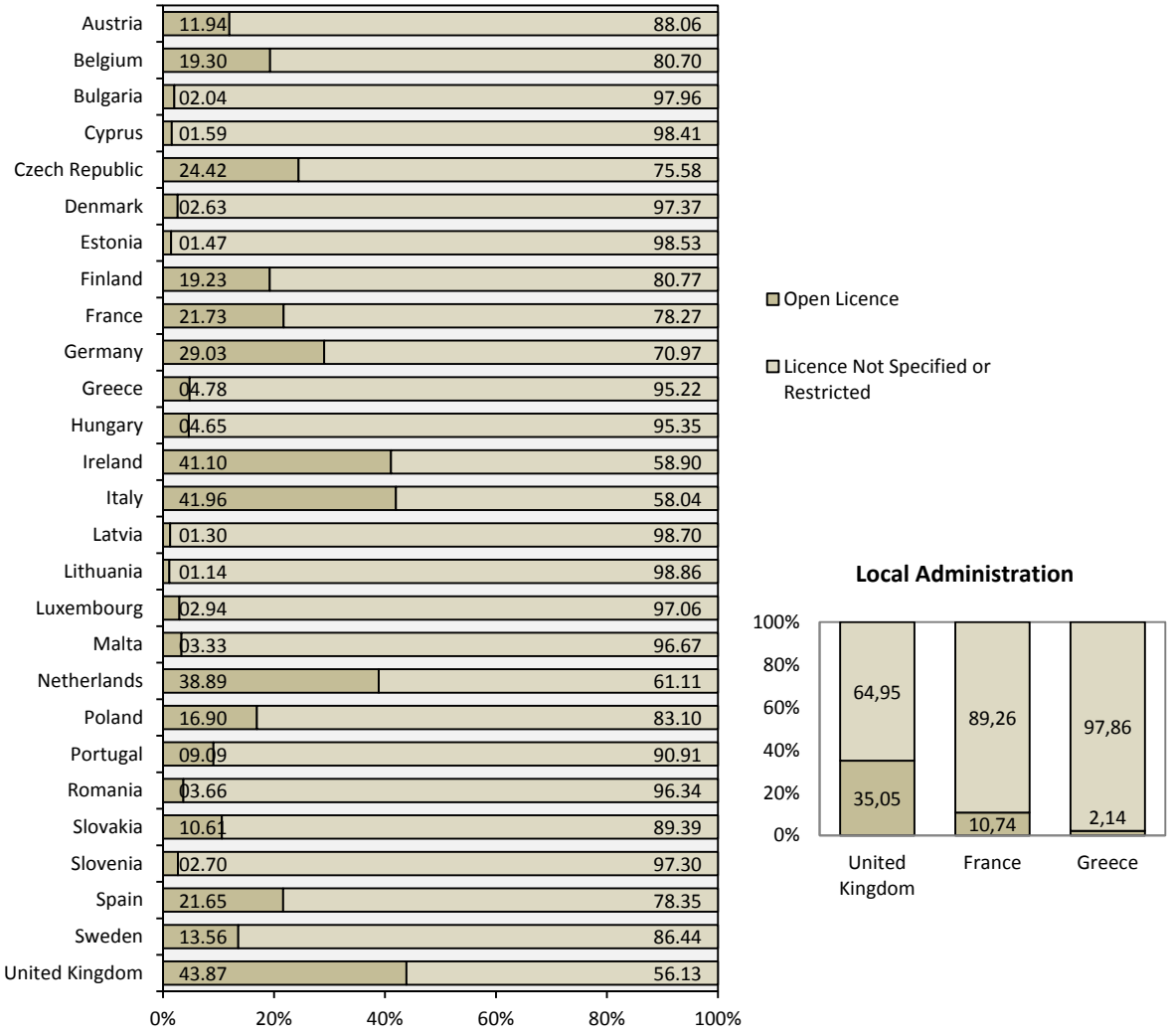



Figure 2: Comparative assessment of the 27 EU countries in terms of their data licensing

Figure 2 (left) illustrates the significant variations in the adoption rates of open licenses amongst the 27 countries. The top four countries with the largest percentage of open licenses are the United Kingdom (43.87%), Italy (41.96%), Ireland (41.10%) and the Netherlands (38.89%). It should be noted that even in those countries the percentage of open licenses is still small, however growing. Moreover, as far as our research is concerned, few open licenses were found in Bulgaria, Denmark, Estonia, Cyprus, Latvia, Lithuania, Slovenia and Malta, meaning that in 8 of the 27 countries publishing government data under an open license was still a necessity. Likewise, by including countries for which the percentage of open licenses was under 5% (Romania 3.66%, Hungary 4.65%, Greece 4.78%) we can assume that in 11 out of the 27, nearly half the EU countries, the application of open licenses needs to be strengthened.

Regarding local administration licensing among the UK, France and Greece, research provided the following results shown in Figure 2 (right). 35.05% of the UK local data sources distribute their datasets with open licenses while 64.95% do it so with restricted or not specified licenses. Similarly, in French local administration 10.74% are published with an open license (89.26% have non-open licenses), followed by Greece where only 2.14% of the datasets were open and the vast 97.86% of them had licenses not specified or restricted. It is apparent that for all three countries the percentage of open licenses by local data providers is significantly lower to that of the national average (43.87%, 21.73% and 4.73% for UK, France and Greece respectively).

## 3.2 Interface and Data Languages

Another point of interest in an ever-diverse and multilingual set of countries such as the European Union is the existence of multilingual support in the data that each provider publishes. More specifically, it was deemed important to separate the case of languages that the user interface, through which the data is provided, is available in from the case of the languages of the data itself. After analyzing the data for both cases, we were led to the following results, shown in Figure 3:

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

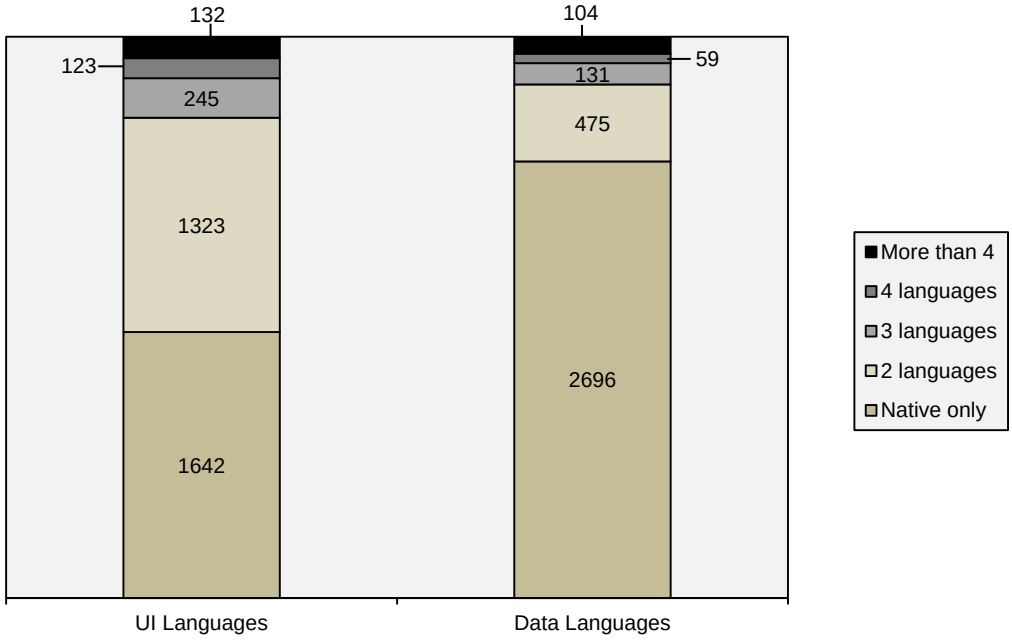



Figure 3: Number of available UI and data languages of the analysed data source

Concerning the UI languages, nearly half of the web interfaces (47.39%) support only the native language of the country they belong to. Additionally, a large percentage of the interfaces (38.18%) are accessible by an additional language, (mainly in English), while 7.07% of them support 3 languages and 7.36% 4 languages or more. The percentages are quite lower with regard to the language of the actual provided datasets. The vast amount (77.80%) of the datasets is available only in their native language. As a result, only 22.19% are available in 2 languages or more, contrary to the UIs where 52.50% of them provide multilingual support. Thus, it is now obvious that the separation of UI and data languages was the only way to showcase the differences between them and to prove that while the former has now reached an acceptable level of multilingual support, the latter remains lacking in that aspect. As expected, the most common language used is English.

In terms of the local administration specifically in the UK, France and Greece that we investigated thoroughly, we were led to the following results. Firstly, UK user interfaces support only one language by 94.85% while 5.13% of them do so for two or more languages. Secondly, 75.12% of French user interfaces are available in French only, with a 24.88% supporting two or more languages. Thirdly, in Greece 57.92% support one language only whereas 42.08% two or more languages. It is evident that in terms of UI multilingual support, local data providers (cities, counties, boroughs, local departments and regions) in France and the UK are further behind than the average 52.50%.

## 3.3 Data Acquisition

This attribute involves utilities that are related to the original source of information as well as the process of retrieving it. Figure 4 summarizes our results per data acquisition method.

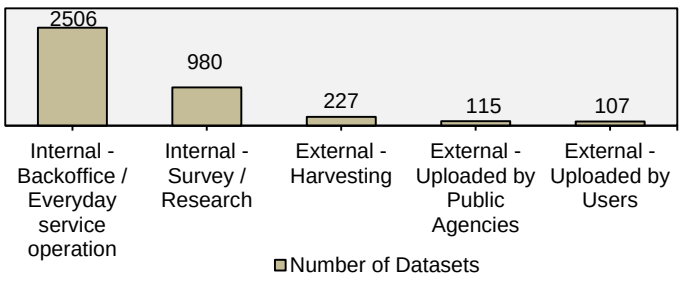



Figure 4: Origin of the acquired EU datasets

As shown in the chart, the majority of the datasets was produced through internal / back office operations of the responsible public agency (2,475 or 72%) and 978 or 28% of the datasets were produced internally for survey and research purposes. On the other hand, the number of the datasets that originated externally was by far lower; 212

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile
This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

datasets or 6% were harvested externally, 115 or 3% were uploaded from other public agencies and, finally, 107 datasets or 3% were uploaded by the users themselves.

It should be noted that often each dataset did not belong only to a certain category; namely, a dataset could contain information that was obtained both internally and externally. In those cases, the dataset was marked in both categories, meaning that the sets of categories portrayed were not independent to each other.

## 3.4 Catalog / Discover

Ease of access to information constitutes one of the most essential features for data providers. Content should be easily accessible and easily discoverable by the users. In this context, after researching the various ways to search and discover data in the providers' websites we reached the following results shown in Figure 5.

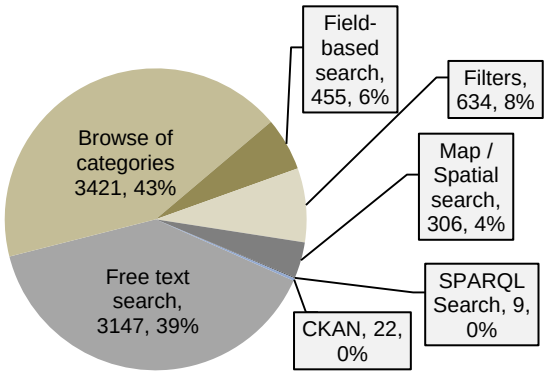



Figure 5: How to access and search datasets in EU governmental websites

The chart illustrates that the two most commonly provided means of data seeking were through common browsing of categories (3386 datasets, or 43%) and free-text search (3112 datasets, or 39%). The rest, more sophisticated ways, follow with far lower figures: use of filters with 8%, field-based search with 6%, map/spatial search with 4% whereas SPARQL search and CKAN have negligible rates.

## 3.5 Data Provision

This attribute refers to utilities that serve the purpose of providing the data / information to users and applications, either in a human readable format or a machine-processable format. For instance, this attribute indicates whether the data is available for online view only, via a downloadable file or both, as well as the existence of value-added services (charts, maps, APIs). The chart in Figure 6 portrays in absolute numbers the different ways the data is provided in EU websites.

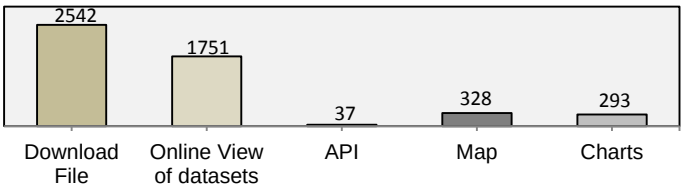



Figure 6: Different ways of data provision in EU datasets

According to the analysis, 2542 datasets are provided in downloadable form (73.34%), 1751 datasets are provided through an online view (50.52%), 328 datasets through a map service (9.46%), 293 datasets through charting capabilities (8.54%) and only 37 through an API (1.07%).

The above analysis was repeated specifically for the national open data portals in order to highlight the difference in data provision between them and the rest of the public data providers. According to this analysis, in 5 out of the 14 at the time available open government data portals, the user can access datasets through an API; a percentage (35.71%) much higher than the average 1.08%. This fact clearly demonstrates the technical maturity of the new open data portals in contrast to the legacy websites of ministries, municipalities and public agencies.

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile
This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

## 3.6 Data Formats

The available data representation formats of the published information is one of the key features of open government portals, as this defines the inherent properties of the datasets, their usability and interoperability. In Figure 7 we portray the most prominent data formats encountered in all of the 27 European countries:

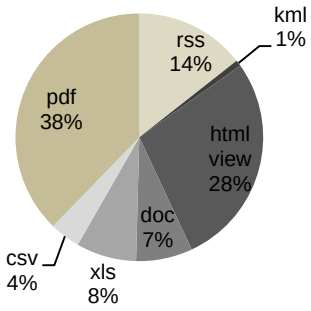



Figure 7: Prominent data formats in EU datasets

The largest percentage of the datasets is stored in the PDF (38%) and in the HTML format (28%). In addition, 14% of the datasets are available through RSS feeds, 8% in the XLS/XLSX formats, 7% in the DOC/DOCX formats, 4% in the CSV format and 1% in the KML format. According to the results we clearly conclude that the majority of the data sources do not provide datasets in machine processable-formats that can be directly consumed through applications, thus limiting their usability significantly.

The analysis was repeated across the EU countries' so as to reach a conclusion about the progress made in each country. The results are shown in Figure 8 (left):

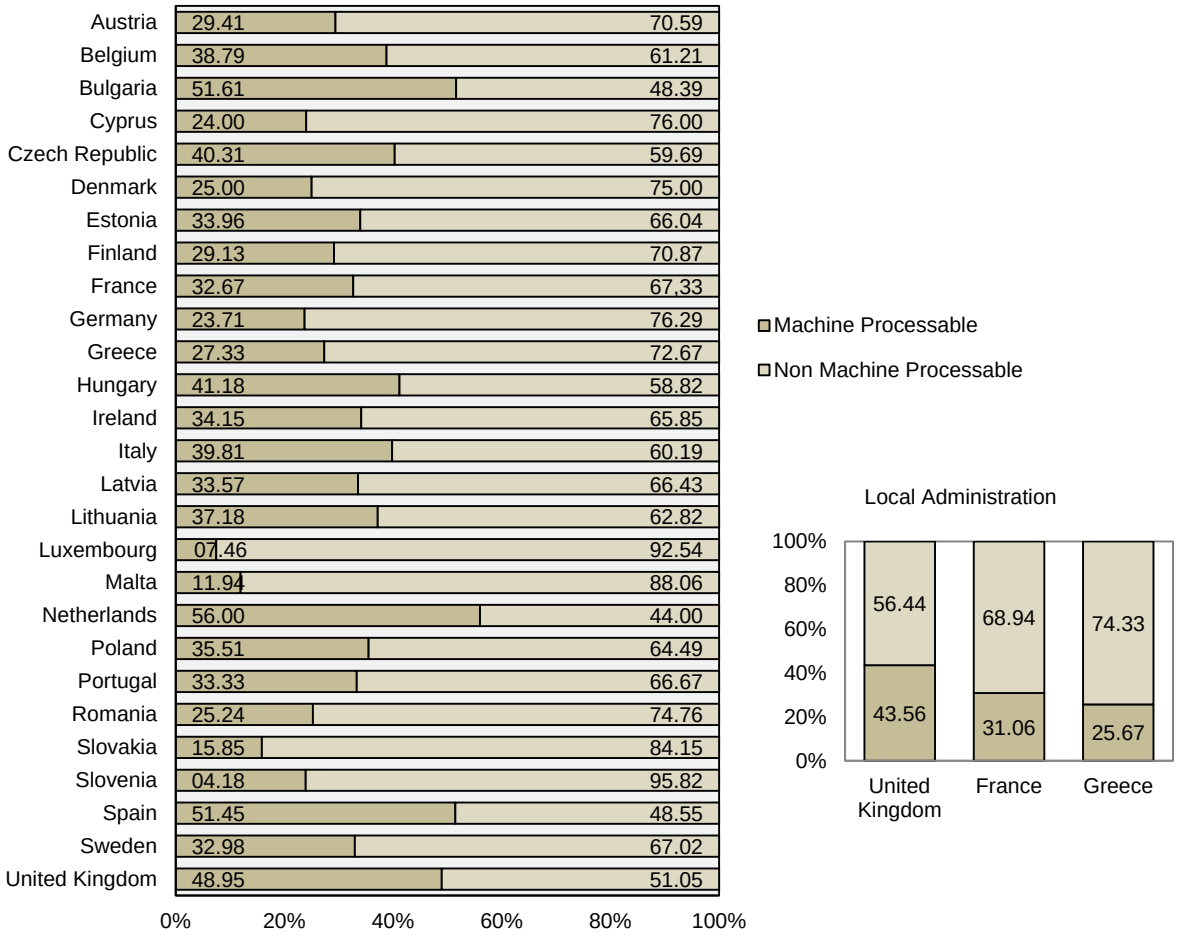



Figure 8: Comparative assessment of the 27 EU countries in terms of the machine processability of their datasets

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

The countries which have made the greatest progress are the Netherlands (56% of the data sources contain machine processable datasets), Bulgaria (51.61%), Spain (51.45%), United Kingdom (48.95%) and Hungary (41.15%). On the contrary, the countries that are lagging the most are Slovenia (4.18%), Luxembourg (7.46%), Malta (11.95%), Slovakia (15.85%), Germany (23.71%) and Cyprus (24%).

Figure 8 (right) also illustrates the findings particularly for local administration data sources in the UK, France and Greece where notably few differences were found compared to their national average. In the UK, 43.56% of the data sources contain machine processable datasets (compared to 48.95% of the UK average) and 56.44% contain non machine processable datasets. Meanwhile, in France 31.06% of the data sources provide machine processable datasets (compared to 32.67% of the French average) while 68.94% do not. Finally, in Greece 25.67% of the datasets are machine processable (compared to 27.33% of the Greek average) and 74.33% are not. Similarly to open licenses and multilingual support, it is apparent that local administration data infrastructure in each country is lacking compared to that of the central administration.

## 3.7 Prominent Category Themes

From a semantic perspective, the number of category themes that each dataset belonged to (Table 1) was also of interest. Amongst the 19 categories, three of them were of particular interest as they turned out to be the ones with by far the largest number of datasets compared to the rest. Firstly, it was the *Law Enforcement, Courts, and Prisons* with 870 datasets (25.10% of the total number). Secondly, it was the *Budget, Revenues, and Expenditures* with 446 datasets (12.87% of the total number). Thirdly, it was the *Business Enterprise, Economics, and Trade* with 336 datasets (9.69% of the total number). As a result, they were particularly studied in terms of their data provision and data formats.

As far as the category of *Law Enforcement, Courts, and Prisons* is concerned, the following results occurred: 754 data sources provide datasets in downloadable form (73.99%), 261 data sources provide an online view (25.61%), 1 data source provides a map service, 1 data source provides charting capabilities and 2 data sources provide an API. According to the analysis, concerning Data Formats in Law Enforcement, Courts, and Prisons, the largest percentage of the datasets is stored in the PDF (53%) and in the HTML format (20%). In addition, 16% of the datasets are available in the DOC/DOCX formats, 10% through RSS feeds, 1% in the XLS/XLSX formats, and a negligible amount in the KML and CSV formats.

The category of Budget, Revenues, and Expenditures it was recorded that 364 data sources provide datasets in downloadable form (60.07%), 173 data sources provide an online view (28.54%), 67 data sources (11.10%) provide charting capabilities, 1 data source provides a map service and 1 data source provides an API interface. As the figures show, regarding Data Formats in Budget, Revenues, and Expenditures, the majority of the datasets is stored in the PDF (44%) and in the HTML format (21%). Additionally, 14% of the datasets are available through RSS feeds, 10% in the XLS/XLSX, 6% in the CSV format, and 5% in DOC/DOCX. KML appears to be of an insignificant percentage.

Finally, Business Enterprise, Economics, and Trade data provision analysis reveals that 248 data sources provide datasets in downloadable form (48.92%), 173 data sources provide an online view (34.1%), 66 data sources provide charting capabilities (13%), 16 data sources provide a map service (3.1%) and 4 data sources provide an API. Similarly, data formats in Business Enterprise, Economics, and Trade are provided in PDF format for the 36%, in HTML for the 26%, through RSS for the 18%, in XLS for the 12%, in DOC/DOCX for the 6.00%, in CSV for 2% and in KML for an insignificant percentage.

## 3.8 Comparative Cross Examination

Below, features of the datasets collected are presented with the intention of their comparative cross examination.

The individual figures in the table above are in line with the average figures already mentioned in sections 3.1-3.6.The overall examination of these figures lead us to the following conclusions:

- The quality of open government infrastructures is steadily improving, and an increasing number of countries try to embrace the PSI directive by developing an official national open data portals.
- There is a great diversity between the national open data portals in terms of openness, interoperability, eInfrastructure and services
- The majority of open data portals and platforms have not tackled multilingual issues. Multilinguality is still a major challenge
- There is no common agreement on the open licenses, each government usually utilizes its own open gov license

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

- Despite some variations, the overall trends that we have presented in our results for the government e-infrastructures of the European Union as a whole are also more or less the same as the ones in each country individually (see Table 2).

Table 2: Comparative cross examination among the 27 countries of the European Union

| Country | Figures |
|---|---|
| Austria | 71.64% of the open data websites are available in 2 or more languages. 79.10% of the datasets are available in downloadable form. The most prevalent file format is the PDF, with a 47% percentage, whereas only 11.94% of the datasets are been published under a free user license. |
| Belgium | 54.39% of datasets are available in downloadable form, hence the obscurity in their utilization and process. 83% of the websites support two or more user interface languages, due to the diversity in Belgium's language patterns (spoken languages). |
| Bulgaria | 85.71% of the datasets are available in downloadable form, while 2.04% of the datasets are accompanied by open license. 84% of the websites support at least one second language, the most prevalent among them being the English Language. |
| Cyprus | The majority of websites (95.24%) are available in a second language (excluding the source, Greek), most frequently in English, but Turkish as well. 73.01% of the datasets are available in downloadable form with a 52% being on PDF format. |
| Czech Republic | 37% of the datasets collected are in English as well. 87% of the websites where the data were collected from support more than one foreign language. 68.97% of the datasets are available in downloadable form. 24.42% of the datasets carry an open license. |
| Denmark | 93% of the datasets have English as a second interface language. 38% of the datasets have data in the English language as well. 65.79% of the datasets are available in downloadable form. |
| Estonia | 67.65% of datasets found are available in downloadable form while 99% of the websites where the data were obtained from support a second foreign user interface language, English being the most prevalent. |
| Finland | 63% of the datasets are in English, while the option of feedback is available in the 67 % of websites. There is no use of open licenses in 80.77% of the datasets, whereas 73.59% are available in downloadable form. |
| France | 73.20% of the language interfaces are solely in French. 71.40% of the datasets are only available online without any storage or further process options. 21.73% of the datasets carry open licenses. |
| Germany | 59.68% of the webpages are available in a second language apart from the source (German). 77.42% of the datasets are available in downloadable form, mostly in PDF format, by 45.36%. 29.03% of the datasets are published under open license. |
| Greece | 72.29% of the datasets found are available in downloadable form. 53% of the websites support a second user interface language. Only 4.78% carry an open license. |
| Hungary | 69.77% of the datasets are available in downloadable form. A 26% is available in English, while in 54% of the websites, where the datasets are collected from, the second language is English. Only in a 4.65% were open licenses found. |
| Ireland | 56.14% of the websites were available in other languages apart from the English language. 98.63% of the datasets are available in downloadable form. 53.66% of the datasets are available in PDF format. 41.10% of the datasets are published with an open license. |
| Italy | 51% of the websites are available in two or more different languages. 82.14% of the datasets were found to be available in downloadable form. 41.96% of the datasets carry an open license. |
| Latvia | 71.43% of the datasets are available in downloadable form. On an 88% of the datasets there was at least a second user interface language. |
| Lithuania | 28% of the datasets are also available in the English language, and 99% of the websites where the data was obtained from carry English as a second user interface language. 73.86% of the datasets are available in downloadable form. |
| Luxembourg | 20.59% of the websites support more than one language. 85.29% of the datasets are available in downloadable form. 72.73% come in PDF format. |
| Malta | 10% of the websites are available in a second language. 85% of the datasets are available in downloadable form (with 70.15% of them being in a PDF format and 8.96% in XLS). 28.33% of the datasets could be viewed online. |
| Netherlands | 42.11% of the websites support more than one language. 73.68% of the datasets are available both for HTML view and for the user to download (download file). The most prevalent format is HTML (26%), followed by PDF format with 18%, CSV with 16% and XLS and RSS with 14% each. Finally, the open license percentage is 38.89%. |
| Poland | 71.83% of the datasets found are available in downloadable form, while in the 83.10% no open license was found. 87% supports at least one second foreign language, the most commonly used one being English. |
| Portugal | 70.45% of the websites are available in more than one language. 84.09% of the datasets are available in downloadable form. PDF is the most prevalent format available with a 50% percentage. 9.09% of the datasets are published under an open license. |
| Romania | 37.80% of the websites are available in more than one language. 92.68% of the datasets are available in downloadable form, and 19.51% for online view. Only 3.66% of the datasets are published under an open license. |
| Slovakia | 72.73% of the websites researched are available in two or more languages. 60.61% of the datasets are available in downloadable form (download file), while 50% can be viewed online (online view of dataset). 54.55% of the datasets are available in PDF format, whereas 50% are also available in HTML view. 10.61% are accompanied by an open license. |
| Slovenia | 86.84% of the datasets are available in downloadable form. 89.47% of the websites under exploration of the websites explored are available in two or more languages. |
| Spain | 77.32% of the datasets are available in downloadable form. 78.35% of data provided are not accompanied by an open license. 81% of the websites support two or more languages, the most prevalent one being English, while in a few cases the alternative options were Catalan Spanish and Basque Spanish. |
| Sweden | 52.54% of the websites under investigation are available in two or more languages. 83.05% of the datasets are in downloadable form, while 47.46% can be viewed online. 71.19% of the datasets is also available in the PDF format, followed by the HTML format with a 35.59% and XLS with 30.51%. 13.56% of the datasets carries an open license. |
| United Kingdom | A high percentage, 43.87% of the datasets, carries an open license. Only 1% of the datasets support a second interface language. 84.19% of the datasets exist in downloadable form. |

**Journal of Theoretical and Applied Electronic Commerce Research**
ISSN 0718–1876 Electronic Version
VOL 9 / ISSUE 2 / MAY 2014 / 34-47
© 2014 Universidad de Talca - Chile
This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

# 4 Application of the Study Results to the ENGAGE Platform (an Infrastructure for Open, Linked Governmental Data Provision towards Research Communities and Citizens)

This state-of-the-art analysis was conducted in the context of the ENGAGE FP7 e-Infrastructures Project (Site 3) and documented in a detailed review of the open data landscape [4]. ENGAGE is a mixture of CP-CSA project funded under the European Commission FP7 Programme and its main scope is to develop and endorse the usage of data infrastructure while incorporating distributed and diverse public sector information (PSI) resources, capable of supporting scientific collaboration and research, particularly for the Social Science and Humanities (SSH) scientific communities, while also empowering the deployment of open administrative data towards citizens. The ENGAGE e-infrastructure is envisioned to promote an extremely synergetic methodology to governance research, by delivering the ground for research to actors from both ICT and non-ICT related disciplines and scientific communities, as well as by ensuring that the methodical results are made available to the citizens, so that they can monitor public service delivery and influence the decision making process.

As of July 2013, in alignment with the present paper, all data contained in this document have been incorporated in the Open Data Section of the ENGAGE platform and this section will be continuously updated until the end of the project covering all 27 countries of the European Union. The Open Data Sites section provides a convenient overview of the Open Data sites landscape analysis as well as a much more efficient maintenance and update of the analyzed open data sites. The Open Data Sites section provides Geospatial search of open datasets, advanced filtering (e.g. per country, category, license, metadata) as well as overall summarized results visualized in bar and pie charts.

# 5 Conclusions

The aim of this paper was to investigate and provide an insight into the current public data infrastructures in the European Union. For this purpose, our research included both central and local government data sources as well as official EU portals and websites in order to ensure that the data sources investigated complement each other and provide the full picture of the current public data landscape. The results of our study show that there is still no uniform policy regarding the provision of public sector information across data sources in the countries of the European Union. The quality of the government data sources varies significantly depending on the country and the data provider. In general, the majority of the datasets is not completely open, as it has been published under restricted or non-specified licenses. Nevertheless, in the recent years there is an increasing effort in the adoption of open licenses, especially in the newly launched national open data portals.

In terms of multilingual support, only 22% of the actual datasets are available in more than one language, whereas in the case of user-interfaces of the data portals (static website text), 52% of them support multiple languages. The discrepancy between the two cases is expected, given the fact that the task of translating the rapidly growing volume of information published by each data provider in more languages other than the original is challenging. Hence, there is a notable difficulty faced by researchers or citizens to access and utilize foreign datasets.

Moreover, most data portals provide the ability to search data only through browsing of categories and simple text search, rarely supporting semantic search – with the bright exception of open government data portals. Thus, the low percentage of SPARQL and CKAN searches as well as the small number of cases where data is provided through an API, clearly indicates that currently there is low usage of Linked Data and Semantic Web technologies. Furthermore, most datasets are published in non-machine processable formats, rendering their technical re- use demanding. This is evident in the case of ministry and other public administration websites, where a simple publication of the data is sufficient, as opposed to the national portals where the availability of technically re-usable formats and semantic interoperability is also a concern.

Despite these shortcomings, it should be noted that the quality of open government infrastructures is steadily improving. Particularly, throughout the EU there is an ever-growing trend of countries, cities and regions towards launching official open data portals where data is published under universally open standards. The United Kingdom was found to be the leading country in that trend, but also France, Austria, Italy, Spain, Germany, Belgium, Portugal, Estonia and the Netherlands have launched their own national and regional open data portals. It is expected that even more countries will adapt to open government policies and the landscape of open government data will vastly advance in the following years.

The Engage platform is a fruitful approach that aims to tackle the variety of the above listed issues. Its architecture and implementation are both focused in a specific goal to understand and adjust into the proposed suggestions, that hopefully more data portals will follow the same paradigm into extended interoperability. Specifically the ENGAGE platform aims to address the challenges that emerged from this study, including the a) multilingual challenge (for instance allowing researchers from Netherlands understand data published by Greek ministries), b) the

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

interoperability issues (by harmonizing the metadata standards followed in each national open data portal), c) Improving the re-use value (data format, content enrichment, data refinement) of open datasets through crowd-based collaboration.

## Acknowledgments

This paper is related to the ENGAGE FP7 Infrastructure Project (An Infrastructure for Open, Linked Governmental Data Provision Towards Research Communities and Citizens; Site 2; Site 3) that started in June 2011. The authors would like to thank their colleagues of the ENGAGE project for their input for this paper although the views expressed are the views of the authors and not necessarily of the project. The results of the analysis in the paper are presented in the open data sites section of the Engage FP7 project (Site 2) with a view to being shared and maintained by the community through user collaboration, thus realizing the vision of an effective public sector information infrastructure [11].

## Websites List

Site 1: Data Governmental Portal of the United Kingdom
http://data.gov.uk/

Site 2: Engage Project Infrastructure Website
http://www.engagedata.eu/

Site 3 Engage Project Website
http://www.engage-project.eu/

Site 4: Greek Ministry of Internal Affairs – Kallicrates Program
http://www.ypes.gr/el/Regions/programma/

Site 5: LinkedGeoData - Spatial Dimension to the Web of Data / Semantic Web.
http://linkedgeodata.org

Site 6: Open GeoSpatial Consortium - Geospatial and Location Standards
http://www.opengeospatial.org/

Site 7: GeoSPARQL - A Geographic Query Language for RDF Data
http://www.geosparql.org/

Site 8: OSMB – Oregon State Marine Board
http://www.oregon.gov/OSMB/Pages/index.aspx

## References

[1] J. C. Bertot, P. T. Jaeger, S. Munson and, T. Glaisyer, Social media technology and government transparency, Computer, vol. 43, no 11, pp. 53-59, 2010.
[2] A. Burton, D. Groenewegen, C. Love, A. Treloar and, R. Wilkinson, Making research data available in Australia, Intelligent Systems, IEEE, vol. 27, no. 3, pp. 40-43, 2012.
[3] Carte de France [online], Available at: http://www.cartesfrance.fr/carte-france-departement/carte-france-departements.html
[4] A. Cordella and F. Iannacci, Information systems in the public sector: The e-Government enactment framework, The Journal of Strategic Information Systems, vol. 19, no. 1, pp. 52–66, 2010.
[5] L. Ding, T. Lebo, J. S. Erickson, D. DiFranzo, G. T. Williams, X. Li, J. Michaelis, A. Graves, J. G. Zheng, Z. Shangguan, J. Flores and, J. A. H. Deborah L. McGuinness, TWC LOGD: A portal for linked open government data ecosystems, Web Semantics: Science, Services and Agents on the World Wide Web, vol. 9, no. 3, p. 325–333, 2011.
[6] European Commission. (2012, October) Directive 2003/9 8/EC of parliament and council on the re-use of public sector information. European Commision. [Online]. Available: http://ec.europa.eu/information_society/policy/psi/docs/pdfs/directive/psi_dire ctive_en.pdf
[7] M. B. Gurstein, Open data: Empowering the empowered or effective data use for everyone?, First Monday, vol. 16, no. 2, pp. 2-7, 2011.
[8] B. Hogge. (2010, May) Transparency accountability initiative, Open data study, 2010. [Online]. Available: http://www.soros.org/initiatives/information/focus/communication/ articles_publications/publications/open-data-study-20100519
[9] J. Hreño, P. Bednár, K. Furdik, and, T. Sabol, Integration of government services using semantic technologies, Journal of theoretical and applied electronic commerce research, vol. 6, no. 1, p. 143-154, 2011.

This paper is available online at
www.jtaer.com
DOI: 10.4067/S0718-18762014000200004

[10] E. K. R. E. Huizingh, Open innovation: State of the art and future perspectives, Technovation, vol 31, no 1, pp. 2-9, 2011.
[11] F. Iannacci, When is an information infrastructure? Investigating the emergence of public sector information infrastructures, European Journal of Information Systems, vol. 19, no. 1, p. 35-48, 2010.
[12] M. Janssen, Y. Charalabidis, G. Kuk and, T. Cresswell, Guest editors' introduction: E-government interoperability, infrastructure and architecture: State-of-the-art and challenges, Journal of Theoretical and Applied Electronic Commerce Research, vol. 6, no.1, vol. I–VIII, 2011.
[13] M. Janssen, Y. Charalabidis and A. Zuiderwijk, Benefits, adoption barriers and myths of open data and open government, Information Systems Management, vol. 29, no. 4, pp. 258-268, 2012.
[14] F. Lampathaki, G. Gionis, S. Koussouris and, D. Askounis, Enabling semantic interoperability in e-Government: A system-based methodological framework for XML schema management at national level, in Proceedings AMCIS 2009 Proceedings, San Francisco, California, 2009, pp. 620.
[15] U. Maier and S. Huber, Open: the changing relation between citizens, public administration, and political authority, JeDEM, vol. 3, no. 2, pp. 182-191, 2011.
[16] F. Mattes. (2012, October) Moving from open innovation to true open innovation, Innovation Management. [Online]. Available: http://www.innovationmanagement.se/2012/10/08/moving-from-open-innovation-to-true-open-innovation/
[17] P. Miller, R. Styles and T. Heath, Open data commons, a license for open data, in Proceedings 1st International Workshop on Linked Data on the Web (LDOW), 17th International World Wide Web Conference, Beijing, China, 2008, pp. 9-16.
[18] S. Mouzakitis, H. Tsavdaris, J. Psarras, J. Klessman, M. Flügge, K. Jeffery, F. Karayiannis and, A. Yaeli. (2011, October) ENGAGE FP7 project, deliverable D7.7.1 analysis report of public sector data and knowledge sources. [Online]. Available: http://www.engage-project.eu/wp/wp-content/plugins/download-monitor/download.php?id=4
[19] Y. Raivio and S. Luukkainen, Mobile networks as a two-sided platform - case open telco, Journal of Theoretical and Applied Electronic Commerce Research, vol. 6, no. 2, pp. 77-89, 2011.
[20] A. Schroll and A. Mild, Open innovation modes and the role of internal R&D: An empirical study on open innovation adoption in Europe, European Journal of Innovation Management, vol. 14, no. 4, pp.475- 495, 2011.
[21] Wikipedia – Subdivisions of England. Wikipedia. [Online]. Available: http://en.wikipedia.org/wiki/Subdivisions_of_England#Hierarchical_list_of_regions.2C_counties_and_districts
[22] T. Zijlstra and K. Janssen. (2013, April) The new PSI directive – as good as it seems?. Open Knowledge Foundation Blog. [Online]. Available: http://blog.okfn.org/2013/04/19/the-new-psi-directive-as-good-as-it-seems/
[23] A. Zuiderwijk, K. Jeffery and M. Janssen, The potential of metadata for linked open data and its value for users and publishers, JeDEM - eJournal of eDemocracy and Open Government, vol.4, no. 2, pp. 222-244, 2012.
[24] A. Zuiderwijk, M. Janssen, S. van den Braak and, Y. Charalabidis, Linking open data: challenges and solutions, in Proceedings of the 13th Annual International Conference on Digital Government Research (dg.o '12). ACM, New York, NY, USA, 2012, pp. 304-305.
[25] A. Zuiderwijk, M. Janssen, and A. Parnia, The complementarity of open data infrastructures: an analysis of functionalities, in Proceedings of the 14th Annual International Conference on Digital Government Research (dg.o '13). ACM, New York, NY, USA, 2013, pp.166-171.