Open Data Portal based on Semantic Web Technologies

Dimitar Trajanov, Martin Kostovski

Open Data Portal based on Semantic Web Technologies

Martin Kostovski[1], Milos Jovanovik[1] and Dimitar Trajanov[1]

[1] Faculty of Computer Science and Engineering,
Rugjer Boshkovik 16, P.O. Box 393, 1000 Skopje, Republic of Macedonia
martin_kostovski@yahoo.com, {milos.jovanovik, dimitar.trajanov}@finki.ukim.mk

Abstract. Recent activities of governments around the world regarding the publication of open government data on the Web, re-introduced the Open Data concept. The concept of Open Data represents the idea that certain data should be freely available to the public, i.e. the citizens, for use, reuse, republishing and redistributing, with little or no restrictions. The goal is to make non-personal data open, so that it can be used for building useful applications which leverage their value, allow insight, provide access to government services and support transparency. These data can contribute to the overall development of the society, by both boosting the ICT business sector and allowing citizens a deeper insight into the work of their government. This recent rise in interest for Open Data introduced the necessity for efficient mechanisms which enable easy publishing, management and consumption of such data. Therefore, we developed an Open Data Portal, with the use of the technologies of the Semantic Web. It allows users to publish, manage and consume data in machine-readable formats, interlink their data with data published elsewhere on the Web, publish applications build on top of the data, and interact with other users.

Keywords: Open Data, Open Government Data, Semantic Web Technologies.

1 Introduction

In the last years, many governments around the world started initiatives for opening up government data which they collect and produce during their everyday activities, and making it available for their citizens to use, analyze and redistribute [1,2,3]. This data is provided in open, machine-readable formats, so it can be used from applications built on top of it, leveraging its value and providing the end-users, i.e. the citizens, with useful information, insight, analysis, e-services, etc. In order to expose this data on the Web and make it available to the end-users, the governments develop and use systems known as open data portals. With the data published on these portals, the citizens can explore the different aspects of the government, as well as contribute in the public decision making process [4,5,6,7]. By publishing government data on open web portals, the governments are giving it back to the citizens, which indirectly paid for their creation with their taxes in the first place. This data creates additional business value within the economy: both individuals and the ICT business sector can

create innovative applications on top of the open government data, and provide services which create new business value and profit [2]. These are the main motivation behind recent research in the fields of Open Data and Linked Data.

Another research field which has been active in the last ten years is the endeavor called the Semantic Web. It represents an upgrade of the current Web, which represents a network of hard-linked documents (web pages with hyperlinks between them), towards a network of information linked up in a way which allows them to be easily processed by machines, on a global scale. This is an efficient way of representing data on the Web, as a globally linked database [8,9]. The Semantic Web aims toward enabling machines to understand the semantics, i.e. the meaning of information on the Web. This will allow the machines to perform more intelligent actions over the vast number of information and data available on the Web today [10]. The technologies of the Semantic Web provide mechanisms for achieving this: they provide a layered structure of frameworks and languages for structuring data into knowledge, by the use of domain specific ontologies [11].

The technologies of the Semantic Web can be used as an integrator across different information, contents and applications for giving more useful data and knowledge, as well as automating many different processes [12,13,14,15,16]. With this, their scope shifts from the Web into many other fields of ICT.

In this paper, we present our open data solution, called the Open Data Portal, which allows users to publish, manage and consume data in machine-readable formats, interlink their data with data published elsewhere on the Web, publish applications build on top of the data, and interact with other users. The application is built with the use of the technologies of the Semantic Web, exploiting their benefits. The Open Data Portal enables the users – both individual programmers and ICT companies – to create and publish their own innovative applications and services based on the data, leveraging the value of the data itself and creating business value.

2 Related Work

2.1 Open Data

Public institutions collect and produce vast amounts of data as part of their daily functioning within the government. This data has a tremendous social value in terms of enabling the citizens to participate in public decision making, to know and have access to government services, etc. Using this open government data, they can contribute to the functioning of government and the democratic structures of society as a whole. A good example would be the mutual communication and trust between the government and society developed through data supply by the government and application building by the society. These types of applications are a good way for supplying the government back with analysis and statistics which they can use to improve their processes and functioning [2]. There is also a huge economic potential within the data, that can be used by businesses which want to add more value to the raw data and thereby take greater market share [17,18,19,20].

As we mentioned before, a great number of countries from around the world started publishing open government data in the last few years. These countries include USA, UK, France, Spain, Germany, Holland, Denmark, Norway, Finland, Italy, Greece and Estonia, as well as Mexico, Canada, Australia, New Zealand and India. Many of them have their own official national data catalogs such as the UK [21], the US [22], Denmark [23], Finland [24], and Spain [25]. Besides these national open data initiatives, there are local initiatives such as those in London [26], Zaragoza [27], Piedmont [28] and Helsinki [29]. The World Bank also has an open data program [30].

Below we present a detailed description and analysis of the Open Data initiatives of the United Kingdom, the United States of America and the World Bank, as one of the most successful open data stories.

UK datasets. The United Kingdom has achieved a significant success in developing its open data portal, http://data.gov.uk/. It features with over 5.600 datasets from all central government organizations and a number of other public bodies and local authorities, which are published in various machine-readable formats, such as XLS, CSV, RDF and XML. The portal also allows downloading the metadata from the appropriate dataset in JSON format. Data can be filtered on several criteria, such as publishers or institutions, nation (England, Wales and Scotland), tags, types and it can also be sorted by relevance, title, ranking, etc. Some particularly interesting datasets available on the UK open data portal include: data about the government ICT projects; trade of goods in the UK analyzed in terms of industry, cost indexes, total number of apartments; data for capital income and expenditure of the Department for Communities and Local Government; consumption in the Intellectual Property Office; data for national public transport; ICT activities of businesses in the UK; weekly fuel prices; index of average earnings; statistics on the labor market; price indices for construction and costs; business investment; monthly statistics on energy; cost data from the Commission on Audit; index services; balance of payments; quarterly energy prices; financial statistics, etc.

US datasets. The US open data portal, http://www.data.gov/, is a leading portal that shows the way for democratization of data and introduces innovation. The data available on the portal is gathered from several locations within the government. The portal has over 305.800 datasets which are published in various machine-readable formats, such as XML, CSV/TXT and RDF. The datasets can be queried by keywords, by categories such as finance, insurance, education, or by public institutions. The datasets include: review of deposits; investment of companies and class information; information about government projects; federal data on business opportunities; foreign labor statistics; index of producer prices; annual data for electricity; business employment dynamics; projections of employees; industry program for funding; resources needed for the budget; consumer and business lending program for investment initiative, productivity and costs; projections for population, etc.

World Bank datasets. The World Bank has an open data portal as well. It has two APIs for providing access to different datasets: one for indicators – time series data – and one for projects – data in relation to the operations of the World Bank. Both APIs implement RESTful interfaces which enable users to perform queries over the available data, using selection parameters. This technology allows for use of the services from both applications and web browsers. The indicators and projects from the datasets are available in XML, JSON and ATOM formats. The indicator API provides access to more than 3.000 indicators, some even spanning 50 years in the past. The API for projects provides access to all of the World Bank projects, including both active and closed projects. Some of the most notable datasets include: World Development Indicators; development finance on a global level; African Development Indicators; indicators about business and enterprises; development goals; global and national education statistics; global and national gender statistics; statistics on health and nutrition, etc.

2.2 Semantic Web and Semantic Web Tools

In recent years, there has been a great progress in the development of tools for representing and consuming knowledge on the Web. These software tools work with the technologies of the Semantic Web and are targeting the processes of design, implementation and maintenance of various semantics and concepts of functioning. Below we describe some of the more advanced semantic web tools based on different technologies, which can be used for developing scalable, reliable and advanced semantic web applications and solutions.

Semantic Web Tools in C# and .NET. DotNetRdf is a semantic web tool based on C# and .NET, falling within the categories of APIs and programming environments [31]. The API supports RDF [32], RDFS [33] and SPARQL as semantic web technologies.

Virtuoso is a semantic web tool used as an RDF repository [34]. Besides the many programming languages available in Virtuoso, such as C, C++, Python, PHP, Java, JavaScript, Action Script, Perl, Obj-C, it also uses C#. It can be defined as a triple store, an RDF generator and a SPARQL endpoint. Virtuoso works with RDF, RDFS, SPARQL, OWL [35], GRDDL [36] and RDFa [37] as semantic web technologies.

RDF Validator is a semantic web tool used as a validator and a visualizing tool [38].

ReDeFer is used as a convertor [39], which can also be classified as a visual RDF generator tool. It works with RDF, RDFa and OWL as semantic web technologies.

RelFinder is used for visualization [40]. It is a tool which extracts and visualizes relationships between objects represented in RDF.

Semantic Web Tools in Java. Jena is a semantic web Java tool, which falls into the categories of triple stores and programming environments [41]. Jena works with RDF, RDFS, SPARQL, OWL and GRDDL.

Protégé is used as an editor and a development environment [42]. It is a semantic web tool that can be classified as a development environment, an editor, a validator and a visual tool which works with RDF, RDFS and OWL.

Another tool that is used in conjunction with Jena is the semantic tool JRDF [43]. It is a tool that is based on RDF and falls into the category of APIs.

Another tool which can be combined with Jena is Virtuoso, mentioned above. The RDF Validator and ReDeFer tools can be used with Jena, as well.

3 Solution Description

We built a platform for open data which aims towards enabling users to easily publish, manage and consume open data in raw, machine-readable formats, interlink the data with other data previously published on the Web, publish applications build on top of the data, and interact with other users. With this, the platform provides the mechanisms which enable innovation in both application and use of open data, by either independent developers or ICT companies. As we pointed out earlier, the software applications built over this open data provide multiple benefits for the society: they provide the citizens with more insight into the work of their government, they create new e-services for the citizens based on preexisting data, they adds new business value into the economy, etc. In this section, we will present the details of our solution.

3.1 Platform Modules

The platform has a modular architecture and consists of six modules:

1. module for displaying the user datasets and publications intended for public use;
2. module for publishing and presentation of applications built by the users, which use the data available on the portal;
3. module for sharing ideas;
4. module for forum discussion;
5. module for blogging;
6. module for displaying the published linked data and the semantically annotated (FOAF) profiles of the users and their relationships;

The first module (Fig. 1) displays the user datasets and publications intended for public use. The common use-case scenario within this module is as follows: the users log on to the system and are given the option to publish their data on the portal. Then they enter additional information in order to meet the input criteria required by the application, and upload one or more files which contain the data which they want to publish. If the files are in an RDF format, then they are submitted for review by the administrators of the portal. If the administrators decide that this data can be published on the portal, they try to find relations between the data and the global linked dataset, and thereby form a knowledgebase. The newly found (if any) relations

and connections of the files are saved in a repository in which the knowledgebase is located, and the files are published on the portal as linked data. This data is then available for other users and visitors of the portal to obtain and reuse.

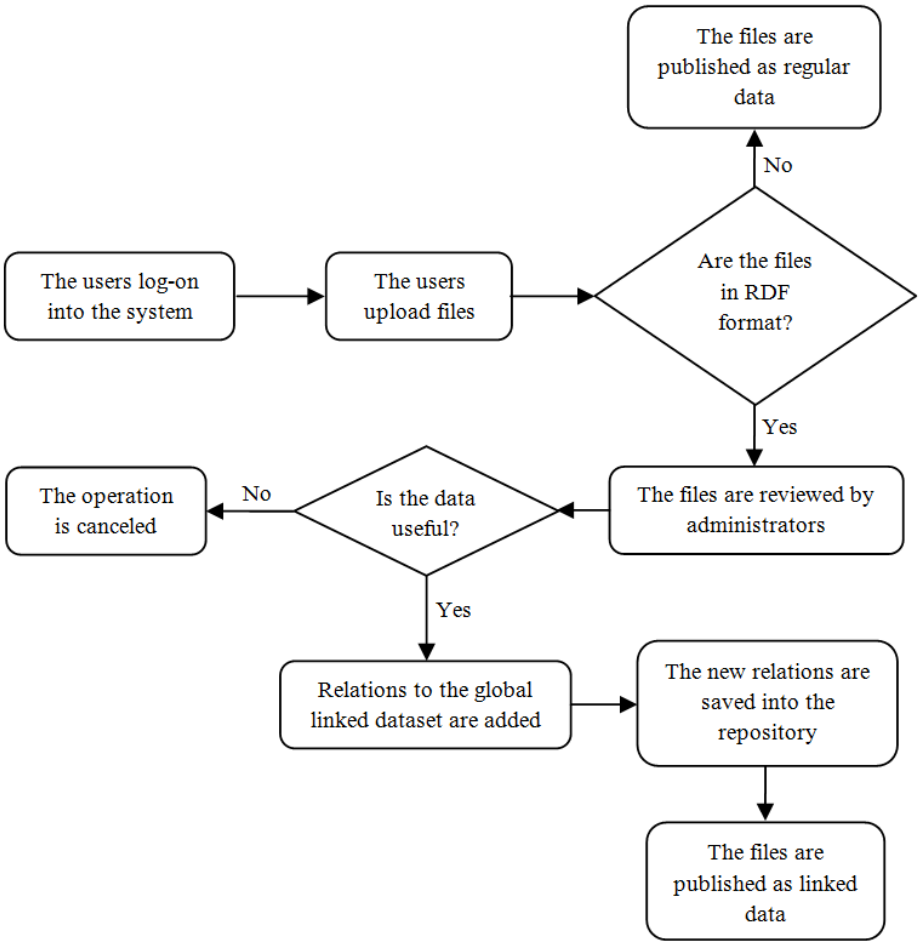

Figure 1: Use-case scenario of Module 1.

The second module (Fig. 2) is the module for publishing and presentation of applications built by the users, which use the open data published on the portal. The common use-case scenario for this module is as follows: first, the logged in users choose to publish a new application. Then they provide additional information to meet the input criteria for publication of the application, and enter the hyperlink pointing to the application, as well as a hyperlink pointing to the open data it uses. After the publication, the application becomes available for other users to browse and use.

The third, fourth and fifth module are places where the users can share ideas, discuss, and publish articles and other content on their blogs within the portal.

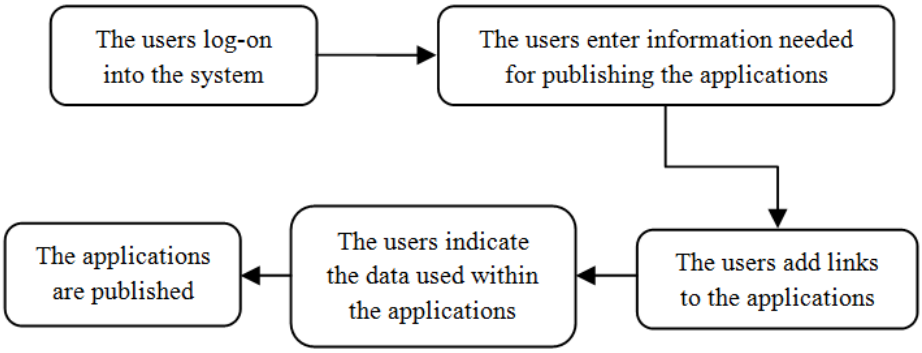

Figure 2: Use-case scenario of Module 2.

The sixth module is the module for displaying the published linked open data, and viewing the profiles of users in a semantically annotated format – FOAF [44]. This module provides a SPARQL endpoint [45] from which users can extract any data from the repository, i.e. the knowledgebase of the portal, by using SPARQL queries. The users can then convert the data into a desired format, such as DOC, XLS, TXT, XML, JSON or CSV. There is also the possibility of visualizing the data from the repository and finding semantic relations between different entities. The common use-case scenario for generating a FOAF profile and creating relations to other people is as follows: the logged in users choose to generate FOAF profiles and connect to other registered users. Then they enter additional information for filling the profiles and enter the links to FOAF profiles to people that they know. After filling in the information, the users generate FOAF profiles and their profiles are published on the portal (Fig. 3).

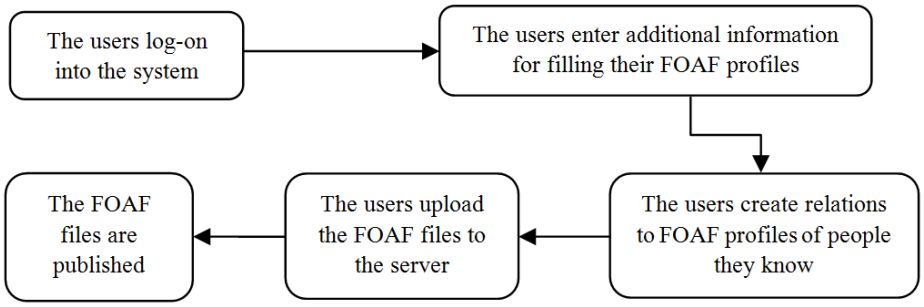

Figure 3: The use-case scenario for generating FOAF profiles of users.

3.2 System Architecture

Our solution consists of a presentation layer, an application layer, a business layer and a data layer, as shown on Fig. 4. The data layer was developed using Microsoft SQL Server, and the other layers were developed using Microsoft ASP.NET technology. The system

contains other elements as well, which are connected to the cloud: a semantic annotator placed within the business layer, an RDF repository represented by Virtuoso and placed within the data layer, and the DotNetRdf API which is placed within the application layer.

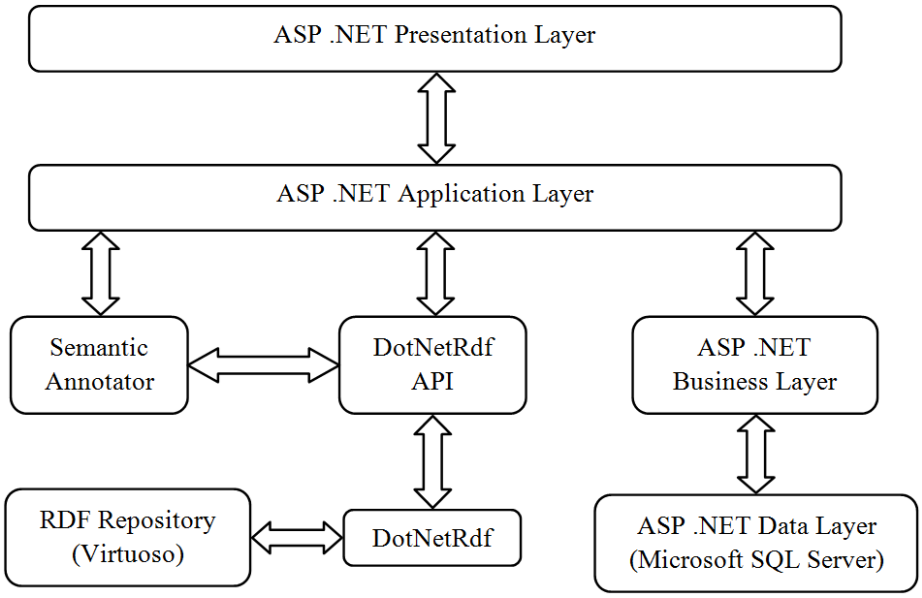

Figure 4: System architecture.

Semantic annotator. The annotator is used to upload RDF files on the server. The content of the file is semantically annotated in the appropriate format and the semantic data is registered in the repository which represents the knowledgebase of the system.

DotNetRdf API. The API provides data access to the repository by using SPARQL queries and data extraction. It extracts and stores RDF triples from the file that is uploaded to the Virtuoso repository. According to the base URI, DotNetRdf sets relations to the global linked dataset, stores them to the Virtuoso repository and updates the knowledgebase based on the relations between the entities.

RDF Repository. The RDF data is stored in the Virtuoso RDF repository as RDF Quads. One RDF Quad consists of a graph IRI, a subject, a predicate and an object within that graph IRI. According to that, RDF triples are taken from Virtuoso and relations are set to the global linked dataset. A knowledgebase is generated from these relations between the entities, and these relations are set in the Virtuoso repository.

Flow of data in the Open Data Portal. The process of data flow within the Open Data Portal for publishing RDF files and generating the knowledgebase is as follows:

- An RDF file is uploaded on the server and the file path is stored within a Microsoft SQL Server database;
- After the file is uploaded, RDF triples from the file on the server are stored in the Virtuoso repository, by using DotNetRDF. The base URI from the RDF file is saved, and according to it, RDF triples will later be selected from Virtuoso;
- After this, the RDF triples are selected from the Virtuoso repository with DotNetRDF and according to the base URI, and relations are set to the global linked dataset.
- If the relations are useful, they are stored in the Virtuoso repository with DotNetRDF, and the knowledgebase is updated, based on the relations between the entities.

The flow of data in the Open Data Portal for publishing RDF files and generating and updating the knowledgebase is shown in Fig. 5.

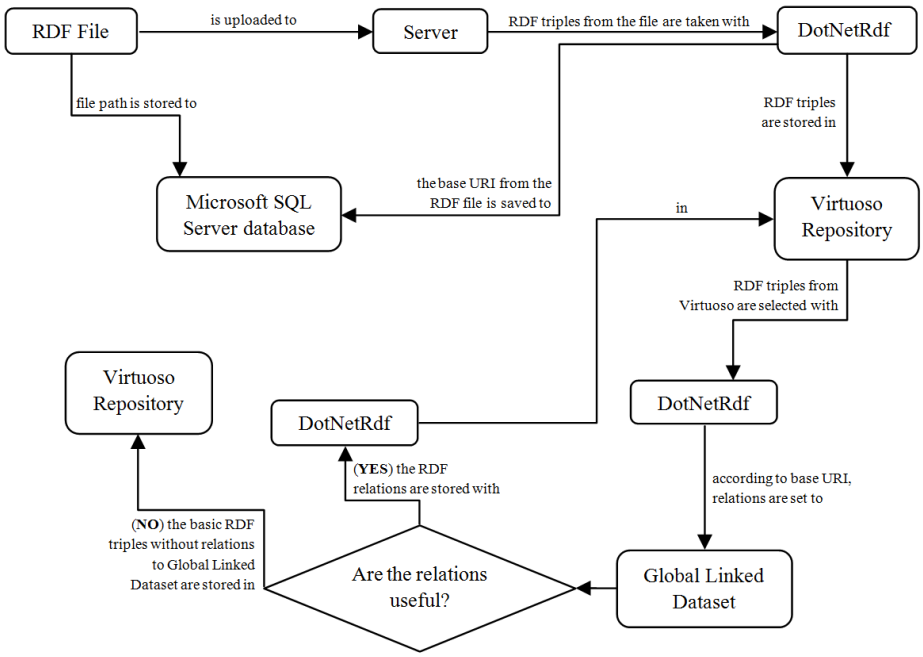

Figure 5: Flow of information in the Open Data Portal.

3.3 Advantages

One of the main motives for the development of the Open Data Portal is enabling the innovative users and businesses to leverage the value of open data, which is provided

by other users, companies or governments. The development of software applications and services over open data provides numerous benefits for the citizens and the society in general, such as greater transparency, communication, acquisition of trust, innovation, and business value.

The basic advantages our platform provides are:

- transparency, innovation and collaboration with citizens and end-users by sharing data, applications and ideas;
- ability for users to highlight and implement their ideas with the help of the open data provided by the Open Data Portal;
- ability to use open data within the business and supply various services to the end-users by developing applications based on the data, and generating new business value;
- ability to interact with other users and join the discussion by blogging and discussing on the forum;
- ability for users to connect with each other by exposing their accounts as FOAF profiles and making connections between them;

The Open Data principle used on this platform brings huge benefits for the end-users and citizens. This system provides a great social for the citizens and the governments, as well as economic potential which can be used by the ICT business sector to build and promote services and applications based on the open data from the Portal. With this, the end-users participate directly or indirectly in the use and consumption of the data, and can use the experience to gather better insight and understanding of the ways their government works. This experience can then be used as a feedback back to the government bodies, in order to provide better services for the citizens. This solution also enables and promotes transparency, innovation and openness to all users and provides assistance in the provision of better solutions and ideas.

3.4 Usage and Performance

The Open Data Portal was tested with 200 registered users. The performance was tested with the YSlow tool [46]. The publishing of data and applications is performed without delay to other processes and the time of publication is a few milliseconds. Finding relations to the global linked dataset does not have major impact on other processes in the system, as well. We tested this process by publishing RDF files with the size of 8MB, finding relations to the global linked dataset and forming/updating the knowledgebase. The problem of finding relations and its effect on system performance is solved through asynchronous communication and execution in the background, in order to reduce the impact on other processes. This is an acceptable solution, given the nature of the problem; we do not need the interlinked connection between the user data and the global linked dataset immediately. Therefore, no performance issues have been detected within the solution.

4 Conclusion and Future Work

The system and the application we designed and developed are intended to demonstrate the advantages the semantic web technologies provide when building an Open Data Portal. This platform should not be viewed as a platform which only offers an end product, but as a place which will constantly be upgraded and hopefully become a base which will offer opportunities for data and application creation, management, sharing and consumption.

Our initiative at the beginning is an attempt to encourage the government to open its doors by opening its data to the public, in order to provide more transparency, participation and collaboration. What we are trying to achieve is to build a strong community by removing the artificial and unfortunate separation of *swimming information*, caused by the fact that different government organizations keep their operating details and data on different online and offline locations.

One of the main objectives in further expansion is the development of applications for mobile devices based on the published open data. With these mobile applications, users will have greater access to services through their mobile devices and will be able to further utilize the benefits of open data.

Our future work will be focused on publication of useful open data, and encouragement of businesses and individuals to develop services and applications which will leverage the value of the published data. One such move would be the development of a module where users and companies will be able not only to publish, but to sell their applications and services. Thus, users and companies will be motivated to produce a great variety of innovative applications and help in the further development of the idea behind the Open Data Portal.

References

1. Berners-Lee, T., Shadbolt, N.: There's gold to be mined from all our data. The Times (2012)
2. Kundra, V.: Digital Fuel of the 21st Century: Innovation through Open Data and the Network Effect. Joan Shorenstein Center on the Press, Politics and Public Policy, Harvard College (2012)
3. Open Government Partnership, http://www.opengovpartnership.org/
4. Ding, L., Difranzo, D., Graves, A., Michaelis, J.R., Li, X., McGuinness, D.L., Hendler, J.: Data-Gov Wiki: Towards Linking Government Data. In: AAAI Spring Symposium: Linked Data Meets Artificial Intelligence (2010)
5. Ding, L., Difranzo, D., Graves, A., Michaelis, J.R., Li, X., McGuinness, D.L., Hendler, J.: TWC Data-Gov Corpus: Incrementally Generating Linked Government Data from Data.gov. In: 19th international conference on World Wide Web (2010)
6. Sheridan, J., Tennison, J.: Linking UK government data. In: WWW Workshop on Linked Data on the Web (2010)
7. Ding, L.; DiFranzo, D.; Magidson, S.; McGuinness, D. L.; and Hendler, J.: The data-gov wiki: a semantic web portal for linked government data. In: International Semantic Web Conference (2009)
8. Palmer, S.B.: 2001-09. The Semantic Web: An Introduction. http://infomesh.net/2001/swintro/

9. Yu, L.: Introduction to the Semantic Web and Semantic Web Services. Taylor & Francis Group, LLC (2007)
10. What is Semantic Web, http://www.altova.com/semantic_web.html
11. Antoniou, G., Harmelen, F.V.: A Semantic Web Primer (Second Edition). Cambridge, Massachusetts, London, England, MIT Press (2008)
12. Nikov, A., Jovanovik, M., Stojanov, R., Petkovski, M., Trajanov, D.: Use of Semantic Web Technologies for Meeting Management Software Development. In: ICT Innovations (2009)
13. Bakraceski, G., Stojanov, R., Jovanovik, M., Trajanov, D.: A Knowledgebase Editor Using Semantic Web Technologies. In: 7th International Conference for Informatics and Information Technology (2010)
14. Apostolski, V., Stojanov, R., Jovanovik, M., Trajanov, D.: Use of Semantic Web technologies in Knowledge Management. In: 7th International Conference for Informatics and Information Technology (2010)
15. Zdraveski, V., Jovanovik, M., Stojanov, R., Trajanov, D.: HDL IP Cores Search Engine based on Semantic Web Technologies. In: ICT Innovations 2010, CCIS 83, pp. 306-315 (2011)
16. Kulev, I., Janevska, E., Jovanovik, M., Stojanov, R., Trajanov, D.: Text Classification Using Semantic Networks. In: 8th International Conference for Informatics and Information Technology (2011)
17. Böhm, C., Naumann, F., Freitag, M., George, S., Höfler, N., Köppelmann, M., Lehmann, C., Mascher, A., Schmidt, T.: Linking Open Government Data: What Journalists Wish They Had Known. In: 6th International Conference on Semantic Systems (2010)
18. Cyganiak, R., Maali, F., Peristeras, V.: Self-Service Linked Government Data with dcat and Gridworks. In: 6th International Conference on Semantic Systems (2010)
19. Berners-Lee, T.: Putting government data online. http://www.w3.org/DesignIssues/GovData.html
20. Alani, H., Dupplaw, D., Sheridan, J., O'Hara, K., Darlington, J., Shadbolt, N., Tullo, C.: Unlocking the potential of public sector information with semantic web technology. In: 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference (2007)
21. Data.gov.uk – Opening up government, http://data.gov.uk/
22. Data.Gov – An Official Web Site of the United States Government, http://www.data.gov/
23. Digitaliser, http://digitaliser.dk
24. Suomi.fi – One address for citizens services, http://www.suomi.fi/suomifi/english/index.html
25. Aporta, Spain's Open Data Portal, http://www.aporta.es
26. London.gov.uk – Greater London Authority, http://london.gov.uk/
27. Zaragoza Open Data Portal, http://www.zaragoza.es
28. Piedmont Open Data Portal, http://dati.piemonte.it
29. Helsinki Open Data Portal, http://www.hel.fi
30. The World Bank – Working for a World Free of Poverty, http://data.worldbank.org/
31. dotNetRDF, An Open Source C# .Net Library for RDF, http://www.dotnetrdf.org/
32. Resource Description Framework, http://www.w3.org/TR/rdf-concepts
33. RDF Vocabulary Description Language 1.0: RDF Schema, http://www.w3.org/TR/rdf-schema/
34. OpenLink Software. OpenLink Virtuoso Universal Server: Documentation, http://docs.openlinksw.com/virtuoso/
35. OWL Web Ontology Language Overview, http://www.w3.org/TR/owl-features/
36. GRDDL, http://www.w3.org/2004/01/rdxh/spec
37. RDFa, http://www.w3.org/TR/xhtml-rdfa-primer/
38. W3C Validation Service. RDF Validator, http://www.w3.org/RDF/Validator/
39. ReDeFer, http://rhizomik.net/html/redefer/
40. Visual Data Web. RelFinder, http://www.visualdataweb.org/relfinder.php

41. Jena – A Semantic Web Framework for Java, http://jena.apache.org/
42. Protégé, http://protege.stanford.edu/
43. JRDF – An RDF Library in Java, http://jrdf.sourceforge.net/
44. The Friend of a Friend (FOAF) project, www.foaf-project.org/
45. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
46. YSlow – Yahoo! Developer Network, http://developer.yahoo.com/yslow/