Publicspending.gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives

M. Vafopoulos, M. Meimaris, A. Papantoniou, I. Anagnostopoulos,
G. Alexiou, I. Avraam, I. Xidias, G. Vafeiadis and V. Loumos

*Multimedia Technology Laboratory, School of Electrical and Computer Engineering,
National Technical University of Athens*

Abstract

The provision of publicly available open data leads to transparency in several public sector exchanges, spending and decisions. However, this information is served massively and heterogeneously - mostly due to different bureaucratic procedures and paperwork formats, while its diffusion does not occur at regular or at least generally predictable time intervals. Thus, even though the information is available by the involved public sectors, enterprises and citizens are overwhelmed from the size/inconsistency of the information they deal with. The scope of our publicly accessible Web point is two-fold. Firstly, it aims to promote clarity and enhance citizen awareness regarding public spending in Greece through easily consumed visualization diagrams. Information provision is based on semantic processing of real-time open data provided by Greek government ("Diavgia") and the Greek Taxation Information System. Secondly, a proposed ontology for public spending in Greece functions in two distinct levels. It checks the validity of the publicly available data accessed by the system, cleaning and reconstructing in parallel false entries, while it will interconnect the data to existing ontological and data schemes derived from other similar initiatives worldwide and core vocabularies.

Keywords: *open data, e-government, public expenditure, Ontologies, citizen empowerment, data journalism*

1. Introduction

The main idea behind the Semantic Web is to extend the Web in a way that information from various sources can be combined independently from applications or content of websites. Moreover, the content itself has to be reorganized in a way that "semantics" hidden behind the information can be interpreted not only by humans, but also by machines [1]. The Semantic Web or Web of Data, as it is lately called, has significantly expanded during the last decade. Its technologies and core tools are now applied to bridge previously autonomous business domains and to concatenate independent government activities [2], [3], [4]. A consequence of the Web of Data underlying technologies brought up lately, the term "Linked Open Data" (LOD) or "Big Data". A term used to describe the combination of data originated from open sources and linked to create a consistent system. In such cases no license terms apply and it is easy and simple to link different data sets to each other. LOD now form a quite extraordinary cloud, but more than that its datasets involve provenance and governmental data (e.g. openspending.org). As Vafopoulos [5] argues "Linked Data enable the creation of better and massive services for use and reuse for many of these data, driving existing infrastructure in its full potential. For government bodies, Linked Data adoption is focused on open, transparent, collaborative and more efficient governance. For enterprises, the core issue is about effective knowledge management and the implementation of new business models that enable more energetic involvement and collaboration between producers and consumers. There is also significant economic potential in Open Government Linked Data, which can be used by business as an input to improve the already existing and create additional value services."

Resource Description Framework (RDF) is the standard data format for dataset resource representation and "RDFizing", in other words transforming all sorts of data to RDF, is being supported by quite a few environments/platforms, like Open Link Virtuoso, Jena Framework, Open Sesame, to name a few.

As far as related approaches with the one proposed in this position paper are concerned, there is a growing number of projects and initiatives worldwide, which target to enhance the citizens' awareness and involvement regarding public sector expenditure. Our approach was mainly inspired by an open-source and embeddable web application entitled "Where does my money go: Showing you where your taxes get spent" (wheredoesmymoneygo.org). Through this application the user can easily find out where UK public finance gets spent by thematic maps and timelines derived from the UK Government open data. A similar project in Europe is the "Open Public Procurement Project" (tender.sme.sk), which in its turn scopes to increase public procurement transparency for the Slovak citizens by cleaning, aggregating and processing procurement data. The web-based system provides complex overview of the procurement processes, thus helping users to compare public spending data as these are disseminated through the Public Procurement Bulletin of the Slovak Office for Public Procurement (e-vestnik.sk). Relevant initiatives for transparency in Asia countries can be found in India and Philippines. The first is called "Accountability Initiative" (accountabilityindia.in) and it is an Indian organization focused on research and creating innovative tools to promote transparency and accountability, mainly regarding government expenditure in public delivery systems. The organization behind the initiative collects data from government websites, where the information is presented in different point of views and formats, organising it into an easy-to-search and sort database. While this initiative uses semi-automatic ways of managing open data, the latter one provides a web 2.0 portal-based dissemination channel that helps Filipino citizens to use text and other rich text format (photos and videos), in order to report occasions where public sector actors are involved in bribery actions. The initiative is supported under the Philippine Public Transparency Reporting Project and it is called "Pera Natin 'to! (It's Our Money!)" (transparencyreporting.net/). Finally, some similar projects that empower citizen involvement and awareness in U.S. are "Texas Transparency" (texastransparency.org/moneygoes) and "Florida Transparency Project: stimulus spending and government accountability" (collinscenter.org/page/stimulus_home). "Texas Transparency" project offers a user-friendly API and search tools that create dynamic reports of an immense amount of data upon user demand.

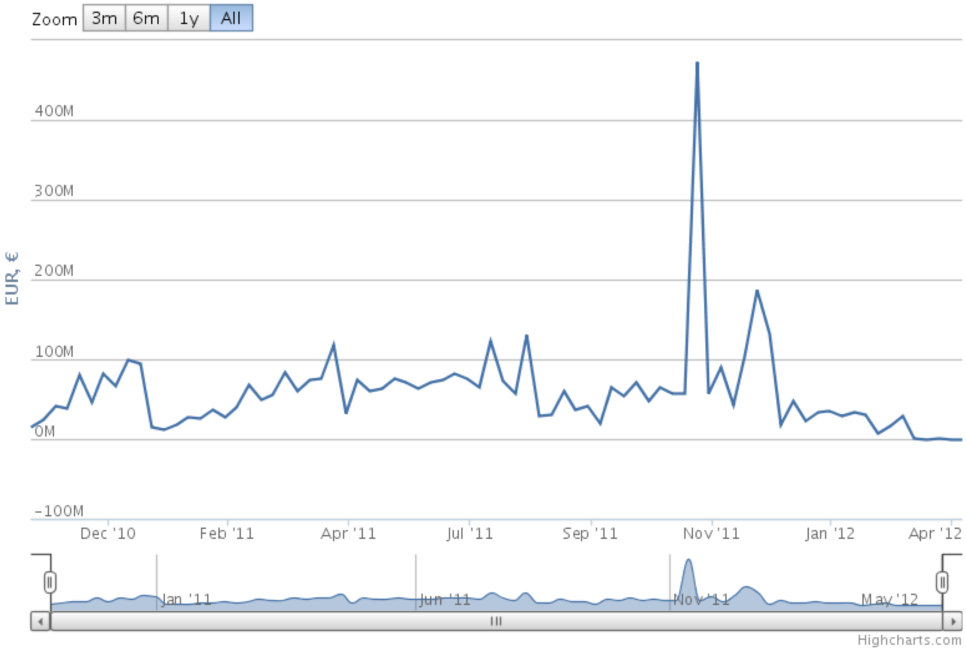

Fig.1: the daily time plot showing the number of spending decisions since the inception of Diavgia.

2. Publicspending.gr: interconnections & visualizations for Greek public spending

2.1 The input data

The present project about the Greek public spending is mainly based on data feeds provided by "Diavgia", the first Greek Government Open Data API (opendata.diavgeia.gov.gr). Diavgia, which is the Greek word for clarity, through its API offers the possibility for publicly accessing all the Spending Decisions of the Greek Public Sector Organizations. Technically speaking, Diavgia is an XML based API, hosting all the public spending decisions in XML files. Metadata include information concerning the types of government expenditures (CPVs), organization types and other (for a detailed description the reader may visit section 3.c). The General Secretariat of Information Systems (GSIS, gsis.gr) is the operating public authority of the Tax Information System (TAXIS), which is the official web portal where citizens and legal entities submit taxation-related information and documents. TAXIS also provides a web service for querying legal entities. The service utilizes the Web Service Definition Language (WSDL) [6], which is an XML format for describing service functionality. The querying is performed in the form of SOAP calls with the entity's VAT registration number as the reference key. The response contains metadata about the legal entity, including contact details, activity descriptions, registration dates and current operational status. Within the scope of this project, the web service is used for querying legal entities on their first appearance as payment agents and the response data are RDFized and stored as payment agent metadata. The evolution of the Diavgia spending decisions, from its kick-off date until now is shown in Figure 1.

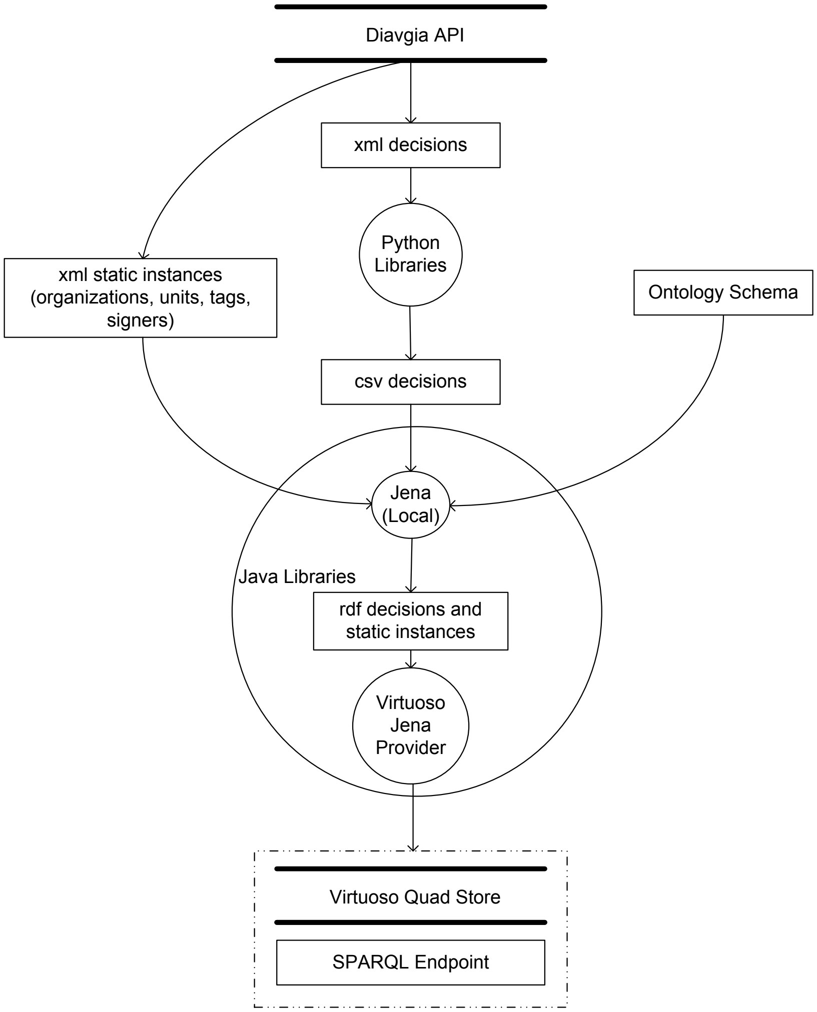

Fig. 2: Project architecture and relevant data flows.

2.2 Architecture and data flow

Currently, the project resides on two Fujitsu Primergy TX100-S3 servers operating on double Xeon processors, deployed as a cluster under Debian Lenny OS with DRBD.

Up to May 2012, the OLV quad store includes almost 2 million decisions, locally hosted in RDF triple notation. The architecture of the project and the relevant data flows are shown in Figure 2. Briefly put, XML decisions are being drawn, on a daily basis, from the Diavgia API, checked for their validity and transformed to CSV format. This process is followed due to the fact that there is a necessity to employ “data cleansing” techniques, concerning both quality (i.e. correctly expressed API terminology) and consistency (i.e. lack of basic fields like VAT, CPV and so on). This “extract and validate” phase was implemented through the use of custom Python libraries. In the second phase, the higher elements of “correct” decisions are further processed, semantically enriched with concepts (resource and property URIs) of the developed ontology (cf. 3.c). This is done by using the Jena framework. The third phase, involves the instantiation of all the RDF decision-related triples (i.e. the actual data) by using core Java libraries, while in the last phase the triples are stored in Open Link Virtuoso environment, simultaneously providing a SPARQL enabled endpoint for data retrieval (diavgeia.medialab.ntua.gr/sparql, username and password can be provided upon request due to current firewall implementation).

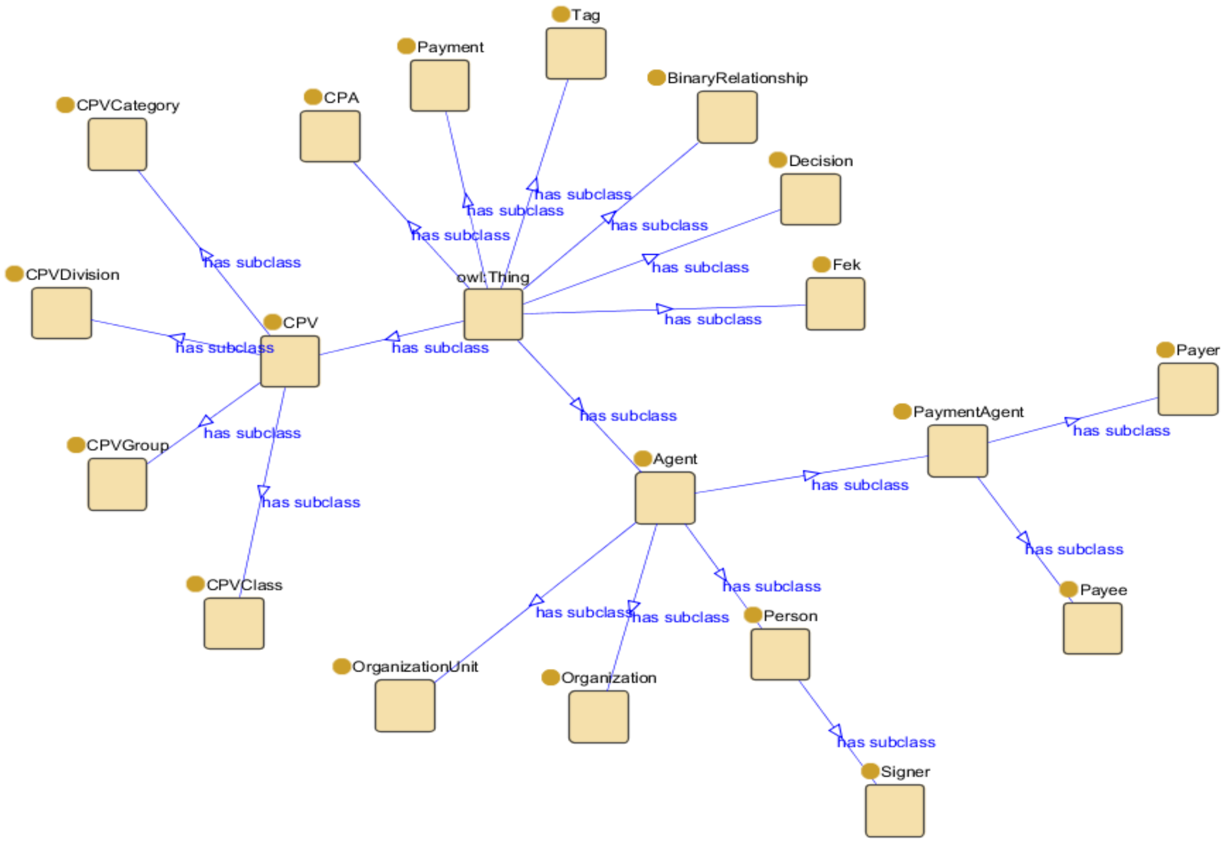

Fig. 3: A visual view of the “Public Spending” ontology.

There is sound reasoning behind the architecture solution and the tools used, which is out of the scope of this position paper. One key issue, though, has to do with the use of multiple and related Semantic Frameworks (i.e. Jena and OLV). The initial intention was to make a completely scalable and framework-free solution and based on the results this was achieved. The project will be officially launched, under the domain publicspending.gr, at the end of May 2012. “Glimpses” of the visualization approaches are illustrated in the subsequent sections.

2.3 The “Public Spending” ontology

“Big Data”, etymologically simply put, involves massive quantities of data that basically need two generic and important processing functions: (a) intelligently fused linkage and (b) visualization. While the latter will be described in brief later on, the first involves the use of “a shared and common terminology”, a phrase that adheres to the notion of ontology. Due to the above, the project presented and discussed within the scope of this position paper couldn’t neglect these widely used shared knowledge models.

The ontology supporting the “RDFizing” of the Governmental Spending Decisions (GSD) API was developed from scratch but it was also inspired by the corresponding ontology of the British “Opening up government” project (data.gov.uk). The basic reasons for our architectural choices in the “public spending” ontology are twofold. First, we followed the “no wheel re-invention”, since the data.gov.uk is considered a widely used initiative. The second factor is related to the fact that it facilitates cross-governmental interlinking between Greece and UK (and other) public spending, which will be implemented in a later phase of our project. The taxonomy part (i.e. sub-classing) of the “Public Spending” ontology is presented in Figure 3. While object and data properties, along with concept restrictions and rules are out of the scopes of this paper, the ontology concept short description is as follows:

- *FEK*: Greek abbreviation, describing the publication of an official government document
- *Agent*: Superclass of agents (people, formal and informal groups)
- *Person*: Class of individual people

- *Decision*: Describes decisions that are published in the Diavgia programme
- *Payment*: Describes payments as unique resources that are referenced by decisions
- *Payment Agent*: Superclass of agents that participate in payments, as either payers or payees
- *Payee*: Describes payees as unique resources
- *Payer*: Describes payers as unique resources
- *Organization*: Official government organizations currently registered with the programme
- *OrganizationUnit*: Operational units of the registered organizations
- *Signer*: Individual officials that have signed published decisions
- *BinaryRelationship*: A resource that describes payment-based relationships between 2 particular payment agents.
- *CPV*: Unique representation of Common Procurement Vocabulary (CPV) codes
- *CPVDivision*: First hierarchical grouping of CPV codes
- *CPVGroup*: Second hierarchical grouping of CPV codes
- *CPVClass*: Third hierarchical grouping of CPV codes
- *CPVCategory*: Fourth hierarchical grouping of CPV codes
- *CPVGround*: Ground level instances of CPV codes
- *Tag*: Represents the list of thematic tags that is supported by Diavgia
- *CPA*: Unique representation of Common Procurement Activity codes

3. Preliminary results

As this work is still in progress, we are one step prior to publicize it. The hosting domain under which it will be run is publicspending.gr. Currently project deliverables run on two “sandbox” domains, one hosting the demo site and one hosting the projects’ SPARQL API endpoint (medialab.ntua.gr/diavgeia/sparql). Through this API one can run SPARQL queries against the dataset and output the results in various formats, one of which being JavaScipt Object Notation (JSON).

The free version of the visualization API of highcharts.com was used in order to visualize the resulting JSON data. Typical examples of the visualization are shown in Figures 4,5 and 6.

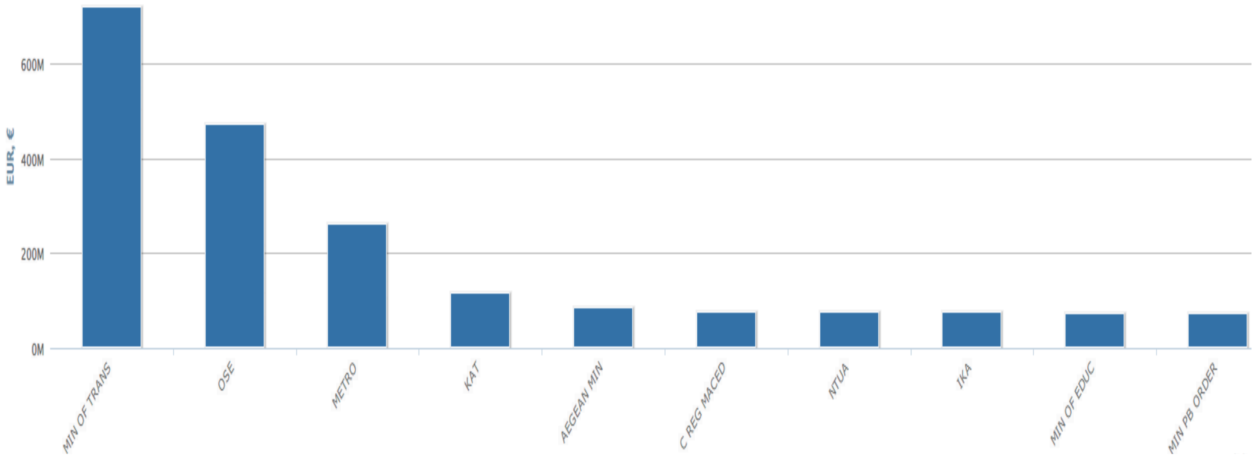

Fig. 4: Greek public bodies with the highest total expenditure from October 2010 to April 2012.

The main distinction between types of graphs is related with the temporal characteristics of the data. Accordingly, four main timeframes, namely daily, weekly, monthly and yearly are considered. Furthermore, the inexistence of time limitations gives one more dimension (overall), which is related with all payments published since the beginning of the Diavgia programme (for example compare Figure 1 with 4 and 5). The second distinction is between static views through one of the aforementioned time frames, and time series views. In the first case, there are “top N” graphs for payments, payers, payees (as in Figure 4) and CPV codes per each of the five time dimensions, whereas in the second case, time graphs are produced for aggregate payments (as in Figure 1) that are associated with the top payers, payees, CPV codes and individual decisions. In the case of time plots, the time period of each node in the graph falls to one of the four time dimensions (excluding the overall view).

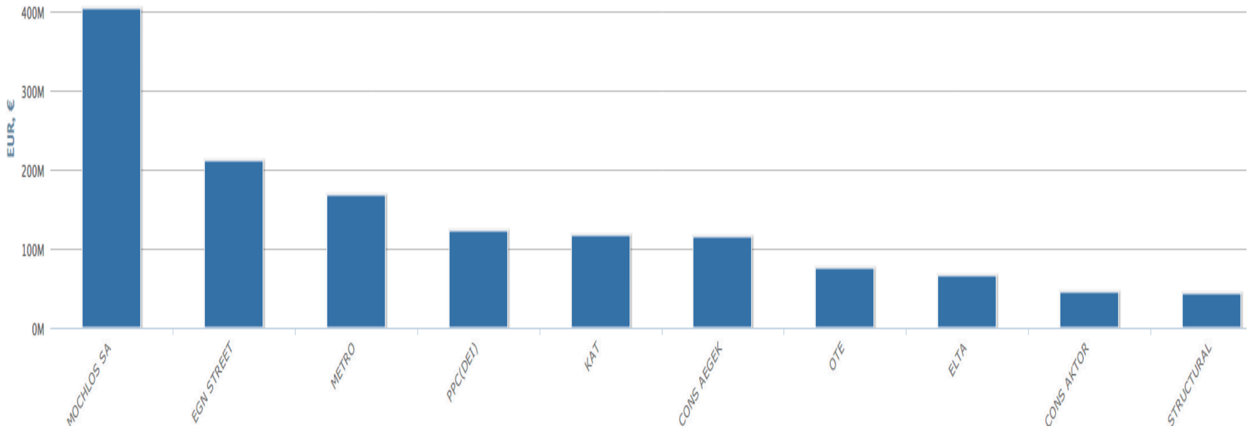

Fig. 5: local and international contractors to the Greek public with the highest total payments from October 2010 to April 2012.

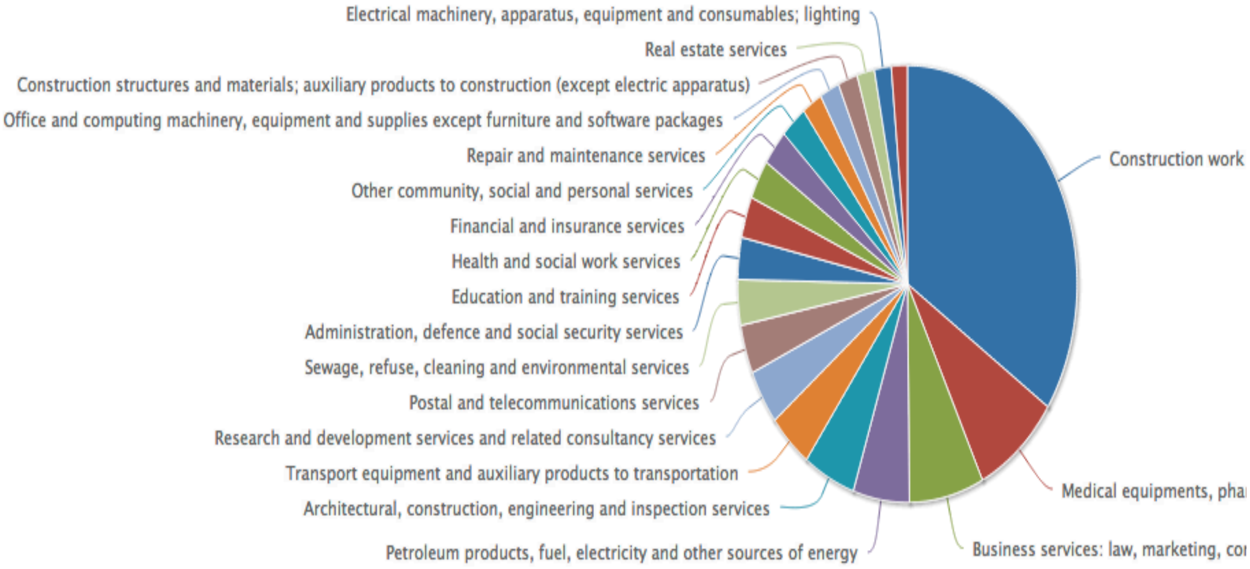

Fig. 6: the top 20 CPV categories of Greek public expenditures from October 2010 to April 2012.

4. Ongoing work and future directions

In our ongoing work we mainly deal with three highly related areas. Firstly, we have to resolve interconnection issues concerning existing "Big Data" datasets. This subject is related to the linkage of our dataset to openspending.org, as well as opencorporates.com and to the core person, business and location vocabularies[1]. To this end, Web Ontology Language (OWL) inference and reasoning constructs are already being involved in our project. Further than these interconnections, emphasis is already being given to the adoption of the Greek Public Contracts Registry (dev.opengov.gr/d/agora/?page_id=322) and to the linkage with the Greek DBpedia (el.wikipedia.org). Secondly, the topic of data and information visualization draws significant attention, especially when dealing with LOD. Apart from the typical graph-based approach described earlier (quite convenient for every Greek citizen), there will be further improvements/enhancements involving Open GeoSpatial APIs (e.g. OpenStreetMap), as well as information cluster maps (e.g. Voronin diagrams).

Last but not least, the development of a concrete and robust API is crucial for our initiative. Since it will be publicly available (including an endpoint) it will become an elaborate linking point to extending and future initiatives, while it will be an innovative tool for social media participants and scientific communities.

[1] http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-1action_en.htm

Future work involves the connection to the Greek National Typography Service (et.gr) and to other developing open data warehouses in local and global scale. Most importantly, will provide a powerful and elastic platform to perform complex and comparative queries in order to identify and highlight incongruities among national and/or European governmental procedures and public sector expenditures (e.g. compare per unit costs of public spending in health, education, etc. among countries and inter-temporally) and to perform planning and evaluation in real time.

References

[1] N. Shadbolt, T. Berners-Lee, and W. Hall, The Semantic Web Revisited. *IEEE Intelligent Systems*, 21, (3), 96-101, 2006.

[2] A. Passant, P. Laublet, J. G. Breslin, S. Decker, “Enhancing Enterprise 2.0 Ecosystems Using Semantic Web and Linked Data Technologies: The SemSLATES Approach”, Linking Enterprise Data, Springer US, 79-102, 2010, doi: 10.1007/978-1-4419-7665-9_5

[3] G. Kobilarov et al., “Media Meets Semantic Web – How the BBC Uses DBpedia and Linked Data to Make Connections”, Proc. European Semantic Web Conference (ESWC2009), Crete, June 2009, doi:10.1007/978-3-642-02121-3_53.

[4] D. Siegel, “Pull—the Power of the Semantic Web to Transform your Business”, Portfolio—Penguin Publishing Group, New York, 2010, ISBN: 9781591842774.

[5] Vafopoulos, M. 2012. The Web economy: goods, users, models and policies. Foundations and Trends® in Web Science (forthcoming).

[6] Christensen, E., Curbera, F., Meredith, G. and Weerawarana., S. Web Services Description Language (WSDL) 1.1. W3C, Note 15, 2001, www.w3.org/TR/wsdl