Insights in global public spending

Michalis Vafopoulos
National Technical University of Athens
9 Heroon Polytechneiou st. 15773,
Zografou Campus, Greece
vaf@aegean.gr

Marios Meimaris
RC "Athena" / IMIS
6 Artemidos St,
Marousi, 15125, Greece
m.meimaris@imis.athena-innovation.gr

Jose María Álvarez Rodríguez
South East European Research Center
24 Proxenou Koromila Street
Thessaloniki, 54622, Greece
jmalvarez@seerc.org

Ioannis Xidias
National Technical University of Athens
9 Heroon Polytechneiou st. 15773,
Zografou Campus, Greece
vaf@aegean.gr

Michael Klonaras
National Technical University of Athens
9 Heroon Polytechneiou st. 15773,
Zografou Campus, Greece
mklonar@gmail.com

Giorgos Vafeiadis
National Technical University of Athens
9 Heroon Polytechneiou st. 15773,
Zografou Campus, Greece
vafeiadis.giorgos@gmail.com

ABSTRACT
Governments around the globe are opening up public spending data in order to promote transparency and citizen awareness. However, data openness by itself is not enough to guarantee that the data is consumed efficiently and in meaningful ways. In this work public spending data from seven governments, both local and national, with total value almost 1,5 trillion euro, are processed, cleansed and converted to Linked Open Data, following best practices. Namely, the cases of Greece, the UK, the US federal government, Australia, the city of Chicago and the states of Alaska and Massachusetts are considered. Furthermore, the resulting Linked Data are interlinked with external resources and made accessible on a public SPARQL endpoint. A web portal application with several functionalities is deployed in order to make the data mashups understandable and easily consumable.

Keywords
Linked Open Data, government spending, network analysis.

1. INTRODUCTION
The Open Government Linked Data initiative aims to improve transparency, accountability and competitiveness in global level. Recent economic analyses show that when re-usable information is provided to the public free cost, then individuals and private enterprises can create innovative and added-value products, which they can then market [12]. Much of this raw data should be curated, interlinked and published in formats that will provide a user-friendly and objective layer of information that will enable citizens, journalists, business people and politicians to re-discover their own data "stories".

In particular, an increasing number of datasets with respect to public expenditure is provided in the form of Open Data[1] and web applications to process and visualize them have been emerged[2]. Practically, all these efforts act as e-catalogs of data that are fragmented in topic, place and time since they do not share common standards and methodologies. In cases where Open Data exist, the basic obstacle is the fact that there are not even standards for representing the main actors (i.e. payers, payees) and the type of payments. Therefore, it is impossible to interlink the available data in meaningful ways and support decision-making.
Surely, the cost of data discovery and collection has substantially decreased, but yet to get the insight in public spending demands high expertise and timely efforts.

This is a serious danger that may undermine the further development of LOD in general. Therefore, we need sound real cases to demonstrate how LOD in public spending can be insightful in making decisions.
We try to overcome the aforementioned obstacles in symbolizing payers, payees and payments by employing transformations and global standards. The scope of the analysis is to demonstrate the power of economic LOD in analysing the market and competition conditions and public policy in global scale.
We found that 480 out of the 2000 companies consisting of the Forbes Global index[3] have received public money from US, UK, Australia, Greece, Alaska, Massachusetts and Chicago. These governments on average distribute their expenditure in construction work (36.7%), office and computing machinery, equipment and supplies except furniture and software packages (15.3%), transport equipment and auxiliary products to transportation (10.3%), financial and insurance services (5.4%) and medical equipments, pharmaceuticals and personal care products (3.6%).

ISEM '13, September 04 - 06 2013, Graz, Austria
 ACM 978-1-4503-1972-0/13/09…$15.00.
http://dx.doi.org/10.1145/2506182.2506192

[1] E.g. www.guardian.co.uk/news/datablog/2012/mar/16/us-open-spending-data
[2] See for instance http://openspending.org
[3] http://www.forbes.com/global2000/list/.

We also employ network analysis; an efficient metaphor to analyze and depict Big and Linked Data because it is intuitively summarizes the mass connections among related entities.

The paper is organised as follows. Section 2 briefly describes the selection of data sources. Section 3 describes the methodologies followed in order to process and triplify the data. Section 4 briefly presents analyses of the resulting networks of resources and Section 5 discusses the implemented application.

2. DATASET

Data selection, apart from the fact that it was restricted by availability, reflects the variety of global economic cycles and policies. Greece[4] is currently under economic distress but leads the efforts to provide real-time information about public expenditure (11). In UK[5] there are concerns about economic slowdown and in USA[6], the source of the 2008 financial crisis, glimpses of hope for development are emerging. On the contrary, Australia[7] enjoys a decade of prosperity. In the local level, the states of Alaska[8] and Massachusetts[9] are not the typical U.S. territories. The first is not considered to be an interconnected and extrovert economy while the second has developed a strong social safety net and a vibrant economy. The city of Chicago[10] is an average big city, which is sensitive in publishing Open Data ranging from the names of child molesters to the salary of each public servant.

3. DATA PROCESSING

The abstract process of generating RDF models from the sources remains the same across datasets, however each implementation has individualities that need to be taken care of in different manners. All aspects of data cleansing, preparation and triplification were done in Java. RDF manipulation was done with the Jena semantic web framework [9]. These will be covered in the following paragraphs, the exception being Greece where public expenditure data already exists in RDF, provided by the publicspending.gr project. For a description of the LD lifecycle that is used in publicspending.gr see [11].

3.1 Data Preparation, Cleansing and Preprocessing

First, each dataset is subject to preprocessing and data-preparation. Even though the data is open, there is no homogeneity in the information the datasets carry, not only across domains (e.g. Greece vs. Australia), but also within the same domain (e.g. data published by different departments in the UK).

Because of the heterogeneity of the descriptions across datasets, minimum sets of characteristics have been identified and universally applied in order to provide a common denomination of the data model. The basic concepts in the public spending domain are the payment and its participants along with their metadata. The minimal set of metadata for each individual payment has been limited to the following: *unique id of the payment, payment amount, payer, payee, CPV subject* and *date of submission*. Optionally upon existence of the relevant data, payments are linked to their decision documents, contract documents and so on. This limit of information is imposed because not all data is useful for our purposes, and different levels of detail would present an information bias. Payments are given URIs based on their unique identifiers (which usually follow different formats) but the URIs themselves indicate the domain of origin.

The use of product scheme classifications is widely accepted in the e-commerce sector to enable the annotation of information objects providing an agile mechanism for performing tasks such as exploration, searching, automatic classification or reasoning. In this context the European Union has established the use of the Common Procurement Vocabulary 2008 (CPV) as mandatory for all public procurement notices according to the Regulation (EC) Nº 2195/2002 of the European Parliament. In order to create a comparison of the contracts object published by different public administrations a hub classification must be used to unify those descriptions. In the present paper the CPV has been selected as a hub classification due to the fact there is a dataset already available as Linked Open Data that links international product scheme classifications [11] to the CPV. Although this dataset contains some exact mappings (created by official organizations), specifically for Greek payments, the rest of them have been created applying custom NLP techniques (making use of the Apache Lucene and Solr APIs) to link a description of a product/service to a CPV concept. Specifically, we have re-used the reported implementation in [13,14] (that is already available as a LOD dataset under the id pscs-catalogue) where authors link together a set of PSCs applying existing NLP algorithms and entity reconciliation techniques. The application of this approach is justified due to the fact that contract descriptions use different product classifications or literals depending on the country, e.g. UNSPSC (Australia), NAICS (USA) or string literals (UK), and they are not available as RDF, so a reconciliation tool such as Silk Server could not be easily used. That is why the existing custom reconciliation service, taking into account the specific characteristics of product descriptions, has been refined in order to link these new product classifications and unify the contracts object with the CPV enabling comparisons among different countries or regions.

Payers and payees are described using some form of unique identification in order to generate URIs, depending on the data origin, as well as their names. The use of different naming techniques such as internal IDs or string literals implies that the task of grouping contracts by a supplier is not a mere process of searching by the same literal. In the particular case of Australia, the supplier name seems to be introduced by typing a string literal without any assistance or auto-complete method. Obviously a variety of errors can be found such as misspelling errors, same company under different names, use of different kind of acronyms [2,3], etc. For example a company such as “Oracle”, “Accenture” or “Capgemini” can be found under different name conventions: “Oracle Aust.”, “Oracle University”, “Oracle Corpoartion (Aust) Pty Ltd”, “Oracle Corp Aust P/L”, “Accentrure”, “Accenture Australia” and “CAP Gemini” among others.

In the Semantic Web area and more specifically in the LOD initiative one of the principles lies in providing a real unique identification of resources through URIs. Thus reconciliation techniques [5,6,7] coming from the ontology mapping and

[4] Data range from 10/2010-11/2012 and retrieved by opendata.diavgeia.gov.gr/.

[5] Data range from 4/2010-12/2012 and retrieved by data.gov.uk .

[6] Data range from 2009-2012 and retrieved by www.usaspending.gov/.

[7] Data range from 2/2006-11/2012 and retrieved by http://data.gov.au/dataset/historical-australian-government-contract-data/

[8] Data range from 7/2007-3/2013 and retrieved by http://doa.alaska.gov/dof/reports/ckbkonline_reports.html.

[9] Data range from 2010-2012 and retrieved by http://opencheckbook.itd.state.ma.us/

[10] Data range from 1/1993-2/2013 and retrieved by http://data.cityofchicago.org

alignment areas or algorithms based on Natural Language Processing have been designed to link similar resources already available in different vocabularies, datasets or databases. The main objective is not just a mere reconciliation process to link to existing resource but to create a unique literal. That is why a context-aware method based on NLP techniques has been designed, customized and implemented trying to exploit the naming convention of the dataset. The technique to generate a unique name before performing the reconciliation process is a stepwise method, in which each step performs a filter over the string literal trying to remove all unnecessary words in the name to finally use an iterative process of string comparison and grouping to generate a unique and relevant name for the input dataset. This process has been implemented using the NLTK library and other external Python APIs such as fuzzywuzzy or a spell checker based on the well-known Peter Norvig speller. The reader is prompted to [15] for more information. After this initial process of unifying names a second step to reconcile names can be easily done reusing resources in OpenCorporates, DBPedia, Linkedin Companies or Google Places. Although a naïve implementation is already available it is considered an extension and ongoing work that is out of the scope of this report. For a particular selection of companies, which was drawn from the Forbes 2000 list, an algorithm has been deployed that aggregates different URIs that represent the same companies in different countries. The Forbes list has been used in order to retrieve objectively interesting companies and create payee supergroup URIs, because of its economic significance. These are in turn linked to their constituents in each country/domain, in order to create points of aggregation for each company. Finally, these supergroup URIs are linked to DBPedia resources manually in order to provide more accurate descriptions and analytics of each company that has received public money, as will be shown in the following sections.

3.2 Triplification and Interlinking

After preparing the data, RDF models were created using the publicspending.gr [11] ontology as well as other widely used vocabularies such as Dublin Core and FOAF. The resulting RDF models were then serialized in RDF/XML form and uploaded to the Virtuoso Quad Store held in publicspending.gr. For each domain of interest the data is stored in a dedicated named graph. This procedure created seven named graphs that cover spending data for Greece, the US, the UK, Australia, the city of Chicago and the states of Alaska and Massachusetts (Table 1).

Table 1: Stats and figures of the seven named graphs

	Triples	Payers	Payees	Payments (€)	#Decisions
AUS	2,788,939	165	51,455	204,606,056,134	429,435
US	494,005	157	16,780	504,599,838,933	61,359
UK	613,3782	119	26,768	597,659,337,423	1,137,641
GR	68,804,546	3904	204,406	48,095,068,098	2,303,360
CHI	63,6526	54	7,930	44,153,540,592	84,405
ALAS	74,8319	22	28,333	16,127,397,046	139,821
MASS	8,023,505	158	31,068	51,279,504,691	1,305,267
Total	87,629,622	4579	315,285	1,466,520,742,920	5,461,288

Furthermore, instances of the same company that operates in different countries/domains have been aggregated under common URIs and linked to their respective DBPedia entries, as was mentioned in the previous section. The named graphs are interlinked in two different levels: (i) on the conceptual model level they use the same vocabularies, and (ii) on the data model through the use of CPV codes and aggregations of companies from different datasets. This makes the superset open to comparisons and analysis across domains, as will be shown in the created application.

4. NETWORK STATISTICS

Despite the fact that network analysis has been successfully implemented in economic networks, it has not yet being applied widely in public spending data. The present section focuses on the use of basic statistics of real networks applied on the representation of public expenditure as a graph. The idea is simple as it relies on the fact that the data produced by payment authorities represent a payment network.

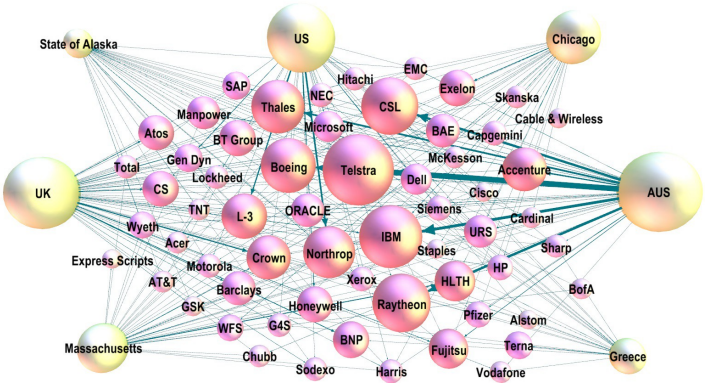


Figure 1: the major vendors of US, UK, Australia, Greece, Alaska, Massachusetts and Chicago[11].

Our objective is to capture interesting relations among underlying agents through graph visualizations. The graph is formulated by the payments coming from public agents (payers) to payees (mainly private but could also be public). A graph node is either the payer or the payee that are linked through a payment, which is characterized by its amount, time and category. A centrality coefficient is a measure that captures the importance of a nodes or link's position in the network. There are local measures like degree centrality (in-out degree, weighted degree) and measures relative to the rest of the network such as betweeness centrality. The main results per dataset can be summarized as follows (see also Figure 1):

United States: there is one major node (payer) in the graph (Department of Defence), dispersing almost all (99%) the total budget (weight) of the graph. Obviously, defence has the lion's stake in sub prime awards. Furthermore, most of the money (92%) is received by CTA Inc., which is solely connected to the department of defence (no connections to other nodes). The dispersion of the public budget is made through 42 agents to the contractors. There are either payer or payee nodes in the graph (no mixed mode both payer and payee-except Smithsonian Inst.), consequently there are no brokers in the network resulting the diameter to equal 1 and the modularity fairly low at 0.022. Due to the above characteristics, there are mainly corporates of the defence-military sector (only US companies are eligible to become vendors due to legal restrictions) coupled by some major global

[11] Visualizations produced by Gephi software.

enterprises with less weighted degree as they do not awarded purely defence contracts.

United Kingdom: is characterized by five major nodes (payers): health, family, education, business innovation & skills, local government disperse 88% of the total budget. The major payees are local authorities or funds responsible for the proper exploitation of the funds received. There are also private companies receiving money for goods and services mainly IT, telecom and consulting. The dispersion of the public budget is made through 26 agents (payers).

Australia: there are two major nodes (payers) in the graph (Department of Defence and Defence Materiel Org.), dispersing almost half of the total budget (weight) of the graph. These two nodes are also the top out-degree nodes in the graph (22% of the payment links). This indicates that Defence is a major factor in the Australian economy sustaining a network of enterprises that selling goods and services suited for the defence needs of the state. The 35% of the budget is spent by institutions related to education, immigration, health and social security, taxation, public order and telecommunication. This reflects, in general, the priorities and major concerns of the Australian state. The dispersion of the public budget is made in a balanced way as there are no private enterprises receiving excessive amounts of money (except FMS and Central Office).

Greece: there are two major projects in Greece: the Subway in Athens and Thessaloniki and Egnatia Runway in North Greece, which involve international companies (e.g. Alstom).

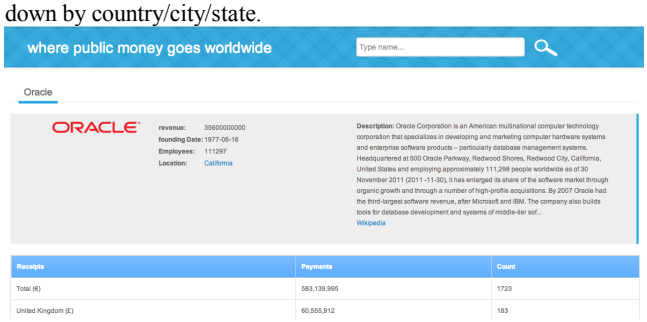
State of Alaska: there are distinct characteristics originating from the special conditions that apply to the region's low population, vast areas of natural resources and ecosystems, weather conditions, native (indigenous) population and distance from global markets. All the above result to a payment network where funds are allocated smoothly to local companies and authorities where health, education, environment, natural resource management, transportation and construction have the lead.

State of Massachusetts: the dispersion of the public budget is made in a balanced way through 157 payers to a network of local institutes and authorities. There are major global players as well but the amounts receiving are smaller due to the bigger amounts that are targeted to health, education and legal institutions and to local authorities. The graph diameter is 1 as there are payer/payee only nodes and modularity is 0.76 indicating the local structure of the payment network. It is worth noticing that there is great variety in the services offered, there are many companies present for every sector (competition) and the balanced value of the in degree indicates a mature market. Massachusetts is famous for its health and educational institutions and this fact is validated from the output data.

City of Chicago is the only city examined (compared to countries or states) but the volume of data ranging from 1993 to 2013 offer a total amount for examination fairly comparable to a state or country. There are distinct characteristics originating from the fact that a city has different needs and priorities from a state/country and of course many resemblances to one (e.g. no need for defence/border safeguarding expenses). The dispersion of the public budget is made in a balanced way through 52 agents (payers) to a network of local companies and authorities and there are also major global players as well. The fact that the city is a transportation hub and resides to lake Michigan is pictured on the graph as major nodes both payers and payees are present and belong to transportation, water management and public utilities sector.

5. THE LINKED DATA APPLICATION

Using the generated RDF triples, an application was built on top of the data in order to serve as a web portal of information retrieval and consumption[12]. Information retrieval can be achieved using the SPARQL endpoint[13] or datahub[14] while information consumption is provided through the use of visualizations and profiling that makes use of internal data, cross-referenced data as well as external data about the RDF resources. Specifically, the application aggregates the functionality that is summarized in the following:
Payee profiles: each payee URI is dereferenced in custom and user-friendly manner in order to combine information about their public receipts and details about the resource itself (drawn from DBPedia). Furthermore, the spending they participate in is broken down by country/city/state.



Picture 1: Profile for Oracle with descriptions from DBPedia and aggregate payment data

Domain profiles: profiles for each of the seven domains (countries, cities and states) are available. These provide detailed descriptions of the places, as well as rankings of their top payees.
Network analysis: statistical analyses for the seven domains are available through the web portal, containing overviews, descriptions as well as quantitative stats about the respective networks of public expenditure.

6. REFERENCES

[1] Bowman, M., Debray, S. K., and Peterson, L. L. 1993. Reasoning about naming systems. *ACM Trans. Program. Lang. Syst.* 15, 5 (Nov. 1993), 795-825. DOI= http://doi.acm.org/10.1145/161468.16147.

[2] Ding, W. and Marchionini, G. 1997. *A Study on Video Browsing Strategies*. Technical Report. University of Maryland at College Park.

[3] Fröhlich, B. and Plate, J. 2000. The cubic mouse: a new device for three-dimensional input. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (The Hague, The Netherlands, April 01 - 06, 2000).

[12] Accessed at www.publicspending.net.
[13] http://publicspending.net/endpoint
[14] http://datahub.io/group/publicspending-net

CHI '00. ACM, New York, NY, 526-531. DOI= http://doi.acm.org/10.1145/332040.332491.

[4] Tavel, P. 2007. *Modeling and Simulation Design*. AK Peters Ltd., Natick, MA.

[5] Sannella, M. J. 1994. *Constraint Satisfaction and Debugging for Interactive User Interfaces*. Doctoral Thesis. UMI Order Number: UMI Order No. GAX95-09398., University of Washington.

[6] Forman, G. 2003. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.* 3 (Mar. 2003), 1289-1305.

[7] Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada, November 02 - 05, 2003). UIST '03. ACM, New York, NY, 1-10. DOI= http://doi.acm.org/10.1145/964696.964697.

[8] Yu, Y. T. and Lau, M. F. 2006. A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions. *J. Syst. Softw.* 79, 5 (May. 2006), 577-590. DOI= http://dx.doi.org/10.1016/j.jss.2005.05.030.

[9] Spector, A. Z. 1989. Achieving application requirements. In *Distributed Systems*, S. Mullender, Ed. ACM Press Frontier Series. ACM, New York, NY, 19-33. DOI= http://doi.acm.org/10.1145/90417.90738.

[10] Carroll, J.J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A. and Wilkinson,K., 2004. Jena: implementing the semantic web recommendations. In Proceedingsof the 13th international World Wide Web conference on Alternate track papers &posters, pp. 74–83.

[11] Vafopoulos, M., Meimaris, M., Anagnostopoulos, I., Papantoniou, A., Xidias, I., Alexiou, G., Vafeiadis, G., Klonaras, M. and Loumos, V., 2013. Public Spending as LOD: The Case of Greece (March 15, 2013). Available at SSRN: http://dx.doi.org/10.2139/ssrn.2233998

[12] Vafopoulos, M., 2012. The Web Economy: Goods, Users, Models, and Policies, in Foundations and Trends® in Web Science: Vol. 3: No 1-2, pp 1-136. http://dx.doi.org/10.1561/1800000015

[13] Álvarez-Rodríguez, J.M., Polo-Paredes, L., Rubiera- Azcona, E., Rodríguez-González, A., and Ordoñez De Pablos, P., 2013. Enhancing the Access to Public Procurement Notices by Promoting Product Scheme Classifications to the Linked Open Data Initiative. *Cases on Open-Linked Data and Semantic Web Applications*, 1, 1--27.

[14] Alvarez-Rodríguez, J.M., Labra-Gayo, J.E., Rodríguez-González, A. and Ordoñez De Pablos, P., 2013. Empowering the access to public procurement opportunities by means of linking controlled vocabularies. A case study of Product Scheme Classifications in the European e-Procurement sector. Journal of Computers in Human Behavior, Special Issue-ICT's for Human Capital. (In press)

[15] Álvarez-Rodríguez, J.M, Ordoñez de Pablos, P., Michail N. Vafopoulos, M., and Labra, J.E, 2013. Towards a Stepwise Method for Unifying and Reconciling Corporate Names in Public Contracts Metadata: The CORFU Technique (July 7, 2013). Available at SSRN: http://ssrn.com/abstract=2290824