Public spending as LOD: the case of Greece

Editor: Jens Lehmann, AKSW, University of Leipzig, Germany; Oscar Corcho, Universidad Politécnica de Madrid, Spain
Solicited review(s): Prateek Jain, IBM Research, USA; Danh Le Phuoc, DERI Galway, Ireland; Hugh Glaser, University of Southampton, UK

Michalis Vafopoulos[a], Marios Meimaris[b], Ioannis Anagnostopoulos[a], Agis Papantoniou[a], Ioannis Xidias[a], Giorgos Alexiou[b], Giorgos Vafeiadis[a], Michalis Klonaras[a] and Vassili Loumos[a]

[a] *School of Electrical and Computer Engineering, National Technical University of Athens, Athens, Greece*
[b] *Institute for the Management of Information Systems, Research Center "Athena", 6 Artemidos st. Marousi, Greece*

Abstract. The Public Spending (PS) project is the first attempt to generate, curate, interlink and distribute daily updated public spending data in LOD formats that can be useful to both expert (i.e. scientists and professionals) and naïve users. The PS ontology is based on the UK payments ontology and reuses, among others, the W3C Registered Organization Vocabulary and the Core Business Vocabulary. RDFized data are linked to product classifications, Geonames and DBpedia resources. Online services contain advanced search features and domain level information (e.g. local government), simple and complex visualizations based on network analysis, linked information about payment entities and SPARQL endpoints. During February 2013, the growing dataset consists of approximately 2 million payment decisions valued 44.5 billion euros forming 65 million triples.

Keywords: public expenditure, Open data, Big data, Registered Organization Vocabulary, economic LOD.

1. Introduction

The Public Spending project (PS) [13] is a research effort introduced by Michalis Vafopoulos to generate, curate, interlink and publish daily updated economic data in LOD formats for both scientists and mass audience. The online service includes advanced search features and domain level data, visualizations based on network analysis, linked data about payment entities and SPARQL endpoint provision. To the best of our knowledge, PS is the first project worldwide to undertake the above tasks to such an extent.

The PS dataset contains information about the public expenditure which originates from the GTI program[1] API and business information provided by official Tax Information System. The RDFized information that is obtained on a daily basis from the aforementioned sources, consists of approximately 2 million payments valued € 44.5 billion that have been paid from 3.900 payers to 204.000 payees and form 63 million triples (Table 1). The dataset is publicly available through the web [2] and a SPARQL endpoint[3]. Data are also available in RDF/XML[4] dumps under the CC-Attribution 3.0 license, given that the original data is also available under the same license.

	#decisions	amount (€ in billions)	#triples (millions)
2010	89K	0.9	2.4
2011	1.031K	12.4	27.9
2012	1.978K	39.3	55.3
2013	2.141K	44.5	63.1

Table 1: agreggated data of PS from October 2010 until February 2013.

The remainder of the paper is as follows. Section 2 is about the source for the data and topic coverage. The purpose of the linked dataset is discussed in

[1] et.diavgeia.gov.gr/en

[2] publicspending.net/greece (in Greek)

[3] publicspending.net/greece/data (in Greek)

[4] datahub.io/dataset/public-spending-in-greece

Section 3. Section 4 describes the applications using the dataset and other metrics of use. Creation, maintenance and update mechanisms as well as policies to ensure sustainability and stability are analysed in Section 5. Section 6 is about domain modeling and use of established vocabularies, while Section 7 refers to the quality, quantity and purpose of links to other datasets. Examples and critical discussion of typical knowledge modeling patterns used are provided in Section 8. Section 9 refers to known shortcomings of the dataset and discusses the way forward.

2. Source for the data and topic coverage

In this section the data sources will be described. The modelling of the domain in a unified ontological schema will be discussed in Section 7.

2.1. The Greek Transparency Initiative (GTI)

Beginning October 1st, 2010, all government institutions are obliged to publish online their decisions with special attention to issues of national security and sensitive personal data. Each document is digitally signed and assigned a unique number certifying that the decision has been uploaded. Furthermore, the decisions cannot be implemented if they are not uploaded on the GTI portal, which introduces an unprecedented level of transparency within all levels of the Greek public administration. Its main objectives concern: (a) the safeguarding of transparency of the governmental actions, (b) the elimination of corruption by exposing it more easily when it takes place, (c) the observance of legality and good administration, (d) the reinforcement of citizens' rights, such as the participation in the Information Society, (e) the enhancement and modernization of existing publication systems of governmental decisions and (f) the availability of all governmental decisions in formats that are easy to access, navigate and comprehend by citizens.

The most important innovation of the program is the combination of closely interrelated legal choices, operational processes and technological strategies. The technological implementation model that is based on an agile strategy of "open content" enacts the dissemination and re-use of Public Sector Information (PSI), providing the necessary tools for thorough access to it. Users are able to "build" applications with added value using the program's content so as to enhance extra functionality. The REST-like Open Data API provides access to all decisions and supplementary information issued by the Greek Public Authorities. The data provisioning is compliant with a specific Transparency Law and are available under the Creative Commons - Attribution license (CC BY 3.0 GR). The latest version of the API[5] supports a taxonomic structure of its content that is returned by default in XML format but JSON, atom and RSS are also available.

2.2. Greek Tax Information System

The second data source is THE Tax Information System (TAXIS)[6] that handles citizens' and legal entities' taxation-related information. TAXIS also provides a web service for querying business entities. The service utilizes the Web Service Definition Language (WSDL) and 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 business entity, including contact details, activity descriptions, registration dates and current operational status. Business activities are described with the use of the Classification of Products by Activity (CPA) vocabulary. Within the scope of PS, 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. In addition, the Greek Tax Authorities during 2012 have started to publish a list of the biggest debtors to the Greek government, including both physical and legal entities.

2.3. Topic coverage

The data from GTI contain detailed information about payers, payees, payment categories, as well as descriptive data related to the spending action (e.g. amount, official decision documents etc. [12]). The data retrieved from TAXIS create an underlying business register layer providing information that uniquely identifies agents as business entities. These include data such as formal names, addresses, legal status descriptions and contact details.

[5] opendata.diavgeia.gov.gr/?lang=en

[6] gsis.gr

3. Purpose of the Linked Dataset

The purpose of the dataset is to enable consumption of the provided data by interested parties, either passively or actively. Passive consumption refers to parties that seek to explore the data in intuitive ways, such as graphs and statistics. With this motivation, a series of daily-updated visualizations are available. We have deployed several applications on the basis of this dataset, the functionality of which is described as follows.

Visualizations: time-series graphs of aggregate payments, bar and pie charts of top 10 measurements of aggregate spending by dimension (e.g. payers, payees and CPV[7]) and by timeframe (day, week, month, year and overall). Advanced spending profiles of organizations and geographical areas.

Search/Advanced Search functionality: users can search through several dimensions such as payment agents, individual payments, geographical areas, signers and so on, and access static descriptions such as business information, as well as dynamically updated information which includes spending data, linked data mashups (e.g. Dbpedia). However, each of these dimensions carries a further and distinct breakdown of characteristics. For instance, payment agents can be searched based on their name, legal form, geographical location, classification of activity and so on, while individual payments can be searched in combination to minimum or maximum payment amount, a range of date and specific procurement category. For this reason, each case was treated separately at the design of the advanced search interface. The relevant queries provide mashups that are in turn visualized in the form of profiles.

Domain level information: focused analysis has been conducted in the spending domains of local government, education, physical persons, companies, health and energy. Statistics and visualizations are presented in the portal for each domain seperately.

Aggregation of spending data: users can see public spending information related to payment agents, (payers/payees), which are registered to operate in the postal code area the user is located.

Crowdsourced payment annotation: the ability to discover and annotate individual payments is also supported. Specifically, when a description of a particular payment is shown, the user can indicate errors in the payment metadata, comment and suggest appropriate corrections. These are usually data entry errors but can also be logical errors, as will be discussed in Section 9. Active consumption refers to usage from developers, researchers, policy makers and professionals in order to create novel results, mashups and applications.

The PS dataset is directly re-used as data feeds in third party websites and as an input to reserarch (see for instance [5] and [6] or the Greek Legal Entities Faceted Browser [8]), data journalism, blogs and political discussions. Innovative visualizations and an important use of PS in the official policy assessment are briefly presented.

Network Analysis: Although network analysis has been extensively implemented in many types of networks, has not yet been applied in public payments. Network modelling [9] is based on the observation that the spending data could be represented as a payment network. The network is formulated by the payments coming from public agents (payers) and are directed to payees (mainly private but could also be public). Nodes are either payers or payees, linked through a payment that is uniquely characterized by its amount, timestamp and category (CPV). According to [15] the public payments network of Greece is connected and sparse. It has 115.734 strongly and 86 weakly connected components. Also, power law has been identified at the indegree distribution. Figure 1 depicts the top payment agents in the Greek spending network.

Policy assessment: a few months after its official launch, the PS project attracted the attention of government authorities. In this context, at the time of the writing of this paper, the operator of the GTI program has already assigned to the PS team various tasks in respect to operations optimization i.e. data quality improvements, usage monitoring reports, statistics on dataset slices and custom reports. The outcome of this bi-directional collaboration has proved to be of important value to both parties. First, the official operator took into account most of our suggestions with regards to the data architecture in the ongoing development of the next version of the program and second, the dataset itself has been evolved to become more stable, structured and functional.

[7] Common Procurement Vocabulary.

[8] http://linkeddata.ihu.edu.gr:8080/rdf-browser/

[9] For a comprehensive introduction to network modelling refer to Section 2.5.1 of 10/3/2014 9:42:00 AM.

4. Creation, maintenance and update mechanisms as well as policies to ensure sustainability and stability

Generation of triples is done using a combination of external data sources, as has been briefly discussed. Each data source is treated separately, as not all types of data exhibit the same degree of dynamicity, the same need for maintenance or the same periodicity of creation. For instance, geopolitical entities, such as regions, do not change and thus their resource representations no not require particular curation, contrary to business registry entities which are updated in an unpredictable way. Decisions are continuously being added to the GTI program, but the backend mechanism of the PS project adds new decisions once a day for load-balancing.

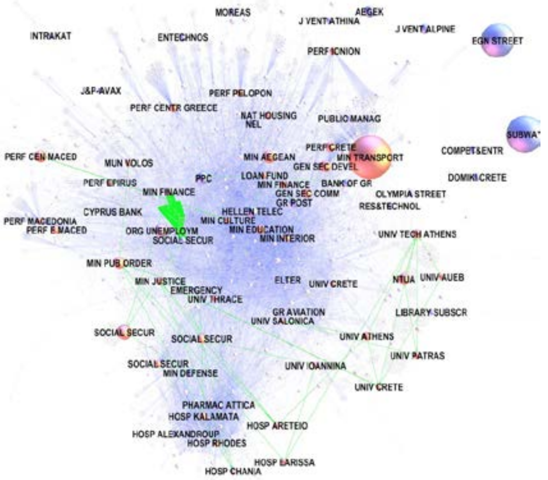

Figure 1: top payment agents in the Greek public spending network from October 2010 until February 2013 (node size: Weighted Degree Centrality and node color: Degree Rank) [15].

Information about businesses is added on first appearance and checked for updates regularly. Ontological schemas and vocabularies are treated and maintained separately. All methods are written in Java and triple models are generated and uploaded using the Jena framework [4]. There are two main types of method invocation for content generation: periodical and on-demand. Periodical generation of Linked Data refers to the process of retrieving and converting data in a timely fashion and following a strict schedule (e.g. new decisions every day), while on-demand generation refers to generating triples from static data whenever appropriate (e.g. maintaining geopolitical resources, CPV/CPA resources, etc.). In the case of routine invocation, requesting, validating and cleaning data, as well as converting to RDF and uploading them to the dedicated triple store is packed in several Java methods that are implemented in a procedural routine way at the beginning of every day in six steps:

1. Decision retrieval. Decisions that are submitted in the GTI program are retrieved using a RESTful API. Users can request full descriptions of decisions based on several search parameters, such as date of submission, issuing organization, signer etc. in XML, JSON, RSS or atom. Not all formats return full descriptions of decisions, and for this reason the created methods handle XML and JSON formats. A mechanism has been deployed that selects between the two, based on current server responsiveness and availability.

2. Response parsing. The responses are parsed and stored in temporary Java class objects that were written to represent individual decisions. Each field value from the response is validated and treated separately during runtime, specifically: (a) string literals are cleaned up from encoding errors and unprintable characters, (b) number literals are formatted and decorated appropriately, (c) errors in number punctuation are identified and corrected and (d) entities are given URIs within the PS domain. URI minting is based on uniqueness, and most of the entities have been assigned unique IDs by the data sources they stem from. These include legal entities, individual payment decisions and so on.

3. Triple model creation. A triple model is created using Jena. Within the model, the retrieved information is given structure with the use of appropriate mapping rules to the PS ontologies and vocabularies.

4. Business registry information retrieval. For each payment agent (either payer or payee) that appears in the triple store for the first time, their VAT registration number is used to form a request, which is in turn sent to TAXIS web service for legal entities and freelancers. This returns as XML, the business information associated with the particular agent. The response is parsed and cleaned-up and the appropriate triples are created and stored in the model. For names and addresses, the string values are transliterated from Greek to English using the ISO 843. The agents are linked to CPA and geographical resource URIs accordingly.

5. Triple upload. The resulting triple model is uploaded to the triple store using the Virtuoso Jena provider methods. These act like a bridge between a local Jena runtime environment and a remote Virtuoso store.

6. Calculation of statistics. At the end of the routine, several measures are derived that are used as references to the dataset's qualitative and quantitative characteristics. These include aggregated amounts of payments by timeframe, number of payers and payees, number of decisions etc. These are stored in a dedicated named graph for statistics.

In the case of generation from static sources, the process is similar to the above, but does not require the same degree of automation, as there is always some control over the process and its execution. The appropriate methods are invoked on demand and the generated data are uploaded after being checked for the consistency in a manual way. Figure 2 illustrates a flow-activity diagram in respect to the processes described above. Figure 3 represents the overall PS architecture of data retrieval, conversion and storage.

5. Domain modeling and use of established vocabularies

Essentially, three distinct ontology schemas have been deployed, as the data were distinguished in three domains, namely spending, organizations and geography. The domain modelling process of each one will be described in brief in the following, mainly focusing on the *spending ontology* as it forms the backbone of the whole model schema.

The PS spending ontology has been designed as an extension of the UK Payments ontology[10], adjusted for the Greek public domain. Public decisions are represented by the *psgr:Decision* core class, instances of which are associated with the decision metadata extracted from GTI, following a similar approach to the UK ontology. Such metadata include dates, subject descriptions, submission timestamps etc. These metadata apply to all different types of decisions. Since payments are not the only decision type available, we adopt an open-world approach on how decision instances are associated with their type. Therefore, while a payment instance holds information on the payment itself, its associated decision object holds metadata that is shared among different types of decisions. This approach aims at extendibility and scalability, given that other types of decisions will be added to the dataset in the future. Payment instances are represented with the *psgr:Payment* class, which contains information that relates a payment with a payer and a payee, as well as a CPV code and a net payment amount. In order to retrieve the payer and the payee associated with a particular payment instance, the *psgr:payer* and *psgr:payee* properties must be used. Furthermore, the *psgr:paymentAmount* property yields the total amount of the payment (in €), typed as an xsd:double literal.

CPV codes are associated with a particular payment instance via the *psgr:cpv* property (Figure 4). It is important to note that since payers, payees and CPV codes refer to resources (as opposed to a payment amount which, in principle, is a literal), querying for these will return the URIs associated with the particular resources, which then have to be further described in order for the information to be human-consumable. The ontology can be accessed online from its IRI. Concerning the organizations, *psgr:PaymentAgent* is the main class in the PS ontology (Figure 5). All payers and payees fall into this class and are identified uniquely by their VAT registration number. However, apart from the agents' unique identification, name and association with payment decisions that are drawn from GTI, we also

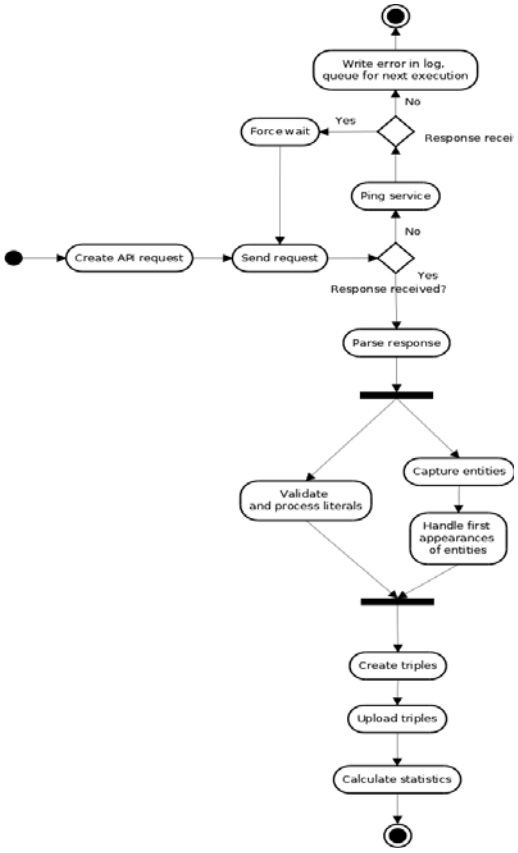

Figure 2: UML activity diagram of the daily routine

[10] data.gov.uk/resources/payments

use external sources (e.g. TAXIS) to draw information that is independent from payments. Among business entities, the main classification factor is their legal form. A lot of work has been done for creating reference vocabularies for business entities, but these are generic as countries use their own legal classifications. This being also the case for PS as all the types of legal entities retrieved from TAXIS are plain string literals.

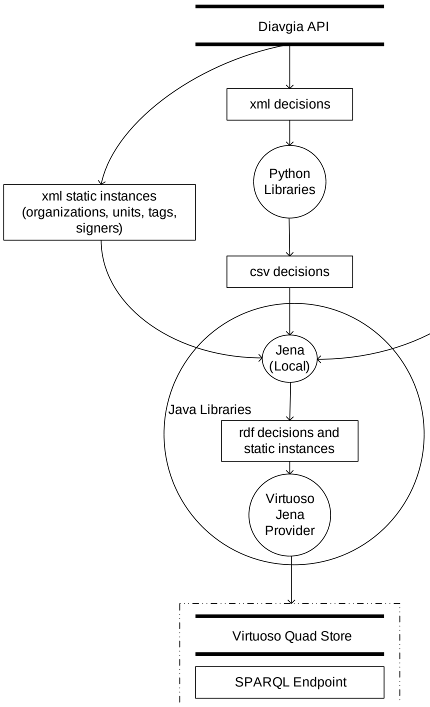

Figure 3: Retrieval, conversion and storing.

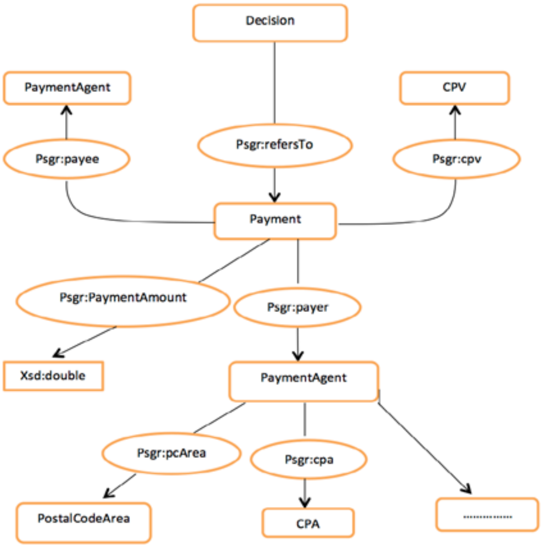

Figure 4: Decisions refer to payments, payers and payees.

To this end an organizations ontology [11] was designed to reflect business entities in Greece also providing mappings with the Core Business Vocabulary (CBV) [12] and the W3C vocabulary for registered organizations [13] at the ontological level in the form of class subsumption. Within PS, business entities are represented as formal organizations, as defined in the W3C vocabulary.

However, as there exists a distinction between physical and non-physical business entities, this issue is covered by differentiating between the two with the use of appropriate classes. Physical entity is no different than the *foaf:Person* class of the FOAF vocabulary [3]. Furthermore, the class LegalEntity is designed as equivalent with Legal Entity from the CBV. The main types of business entities have been modeled for Greece by their English labels. Mappings from literals to classes were performed with the use of Jena and some basic NLP methods on the string descriptions of the legal forms. Within the PS dataset each payment agent is associated with a postal code. Given that, we have provided a class, *psgrGeo:PostalCodeArea*, to describe areas defined by postal codes. Postal code areas belong to broader areas which can be municipalities and regions. This pattern created the need for an ontology to efficiently describe it. The *psgrGeo:isPartOf/hasPart* property connects PS ontology classes. Each class provides a generic set of properties in order to represent the area that it refers to. These sets of properties include names, postal codes, LAU/NUTS [14] level, latitude and longitude. Since the ontology only describes the structure that is mentioned above, the dataset is connected by the *owl:sameAs* property with DBpedia [2] and Geonames in order to further extend it, as will be discussed in the following section.

6. Quality, quantity and purpose of links to other datasets

As it has been mentioned, each payment is associated with a particular CPV and each business agent is assigned a unique CPA by the Greek taxation authority, which is available in the PS dataset.

[11] publicspending.net/organizationsOntology.owl
[12] joinup.ec.europa.eu/asset/core_business/description
[13] w3.org/TR/vocab-org/
[14] geovocab.org

However, procurement classifications do not follow a unique standard. For instance, the European Union uses the CPV, while the USA follows NAICS[15]. The MOLDEAS project [1] is an effort to align this plethora of classification systems and publish them as RDF with controlled semantics, following SKOS [8] modelling principles.

URIs representing CPV and CPA codes in our dataset are linked (with *owl:sameAs* links) to their corresponding codes within the MOLDEAS dataset, paving the way for comparisons on different classifications of spending (Table 2). Recently, [10] introduced a stepwise method based on NLP and semantics to address the unfication of corporate names. The second area of interlinking refers to geopolitical information that is associated with business register entries. Given that each business agent is associated with geospatial information (in the form of associations with postal code areas), we have identified and linked the relevant geopolitical resources (regions, regional units, municipalities, country) with their representations in DBpedia and Geonames (Table 2).

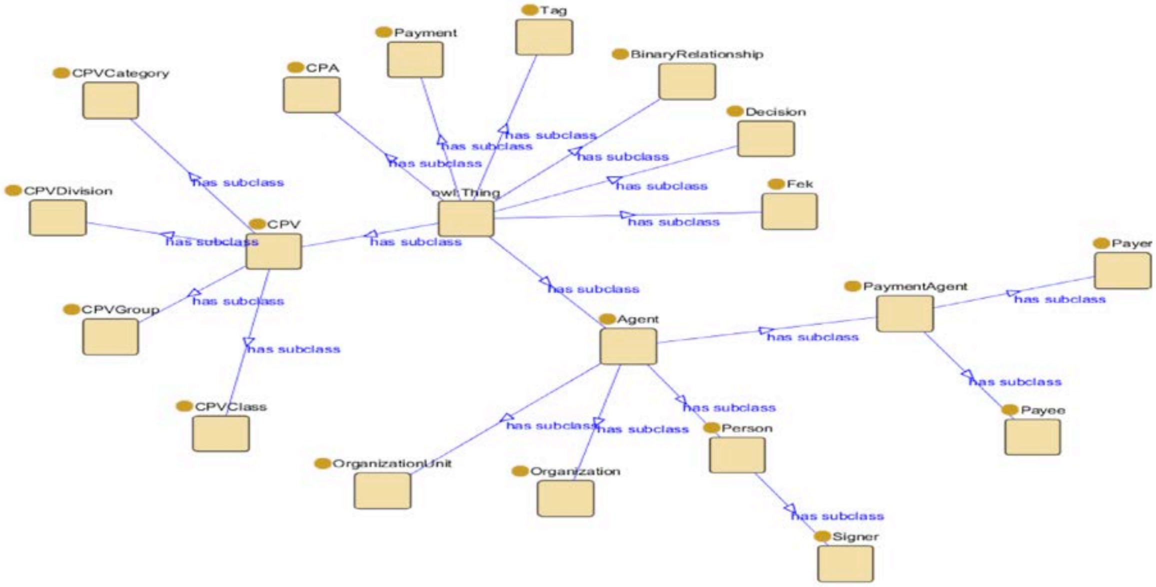

Figure 5: PS spending ontology has been designed as an extension of the UK Payments ontology adjusted for the Greek public domain.

	MOLDEAS	Geonames	DBpedia
number	9.454	402	804
purpose	product classifications	geonames resources	payment agents

Table 2: number and purpose of PS interconnections (up to February 2013).

In particular, we performed the mapping between postal codes and postal code areas (i.e. geopolitical areas uniquely defined by postal codes in a 1:1 fashion) in order to create postal code area resources and, in turn, link to geopolitical resources at different hierarchical levels. This enables the association of spending and business register data with resources that exist within DBpedia and Geonames, so the formation of queries such as finding statistical information of spending by area, or spending by population, becomes trivial. Furthermore, as we traverse the DBpedia graph we can associate the PS dataset with a large amount of external resource nodes with a linking path connecting them semantically, giving answers to potentially interesting queries. Currently, we are building direct connections of the payment agents to Greek, English and Live DBpedia [9]. The linking is done in the geoinformation level since we decided not to link the authority, but the geopolitical region that the authority serves (e.g. distinction between the authority “Municipality of Athens” and the region that is governed by this authority). This procedure has been completed for local government authorities and is ongoing for the payees coming from the private sector. Already, we have identified more than 1000 private agents having a canonical name. In addition, we initiated the linking with data related to

[15] census.gov/eos/www/naics/

the Greek economy. Local Tax Authorities during 2012 have started to publish an updated list of the biggest debtors to the government, including both physical and legal entities. However, the data was published as plain HTML with no actual URIs or web resource identifiers of any sort. For this reason, we had to mint our own URIs and generate triples to represent this information in our dataset. A cross search between payees and debtors of the public concluded that 416 entities are both payees and debtors to the public during February 2013.

7. Examples and critical discussion of typical knowledge modeling patterns used

Being "pragmatic" within the context of the aforementioned methodologies, in other words following simple approaches as compared to "heavy" ones such as commonKADS [11], was the chosen path. Adopting "heavy" ontology engineering within LOD initiatives, not only hinders Web Scientists [7,14] to provide mere outcomes but affects the performance and flexibility of underlying schemas. However, a strong motivation for economic LOD is for their respective datasets to be able to support interoperability, not just between authorities in the same country, but at an international level as well, so that comparisons can be feasible and meaningful. This means that the established vocabularies and data should be structured in ways that are descriptive enough to support the linkages. For instance, there is no unambiguous 1:1 matching of legal entity types between countries. The Registered Organization Vocabulary [16] is a higher level vocabulary, but perhaps too high for this purpose, as its expressiveness is limited to representing organizations that are recognized by some official authority, and registered within it. An interesting research topic for future work is to create a mapping schema of close matches between different types of legal entities, or a vocabulary that aggregates most commonly used close-matching legal entity types from different countries under higher-level superclasses (e.g. a generic class to represent all company types that bear close resemblance, such as PLC, SA). Without this step, there is no direct way of bringing companies from different countries down to a common denominator and compare.

8. Known shortcomings of the dataset and future directions

Since the first initiation of the GTI project more than 30K public organizations in Greece use the web application to upload every single decision, which is digitally signed. Throughout the investigation we noticed that there are several errors derived mainly during the data entry phase, due to the lack of appropriate restrictions and validation mechanisms. Serious limitations have to do with *data format errors* (e.g. numbers as strings), *incorrect entries* (e.g. VAT or payee name) and *missing values* (mainly CPV codes).

Having identified the major data format errors, we have developed a series of automatic error correction mechanisms. As a result of this, nearly 200K and 4K decisions with erroneous VAT entries and CPV entries have been identified, respectively. These corrections reflect a total amount of expenditures that reaches the level of 10 billion euros.

Syntactic or semantic types of errors mainly drive incorrect entries. Syntactic errors on the VAT or payee name are corrected by calling the TAXIS service. Semantic errors are much harder to handle since they involve manual detection and correction. Since the launching of PS, it was noticed that for several physical entities (individuals), the amount of received payments was enormous and in the majority of the cases unreasonable. A closer investigation revealed a significant problem, which arose due to an erroneous procedure that some users follow during the data entry. To be more specific, let us consider the case where a social security authority (payee) receives X euros to provide the welfare allowance for Y beneficiaries. Each one of the Y beneficiaries has a distinct VAT registration number. Nevertheless, in most of the cases in the GTI program, a sole physical entity such as the legal representative of the payee or even in some specific cases a random representative out of all beneficiaries (usually the person whose name apperars first in alphabetical order) is stated erroneously as the corresponding payee. This procedure occurs as a "solution", mainly because there is no multiple VAT registration number provision to complex payments. As a result, some physical persons along with their VAT registration numbers are "loaded" with huge payments, which are totally irrelevant to their income. Trying to provide a solution, we followed a semi-automatic procedure that helped us clean the respective data. Through SPARQL queries we gathered the top 2000 physical

[16] w3.org/TR/2013/NOTE-vocab-regorg-20130801

persons (according to their VAT registration number) who were labelled as payees from the beginning of the GTI program up to the end of September 2012. Then for each separate record, we reviewed and cleaned their top 10 payments. The total amount of updated/cleaned payments reached the level of 550M euros.

Data cleansing was only performed to the extent that the meaning and intention of the data was not compromised. In any case, we track provenance by marking up decisions as *complete*, *faulty* or *corrected*. In the case of faulty and corrected decisions, we include the information of what was faulty and/or fixed. Apart from providing the users with digested financial information regarding public organizations (which until now was only estimated by speculation and sometimes sketchy data), the dataset and the applications provide a direct view of the quality of the original data uploaded by each organization. For instance, there are organizations that consistently abuse the upload form and provide incomplete ορ erroneous data. These findings have been reported to the official authorities who in turn took action to further discipline the organizations and upgrade the application. Interestingly enough, we have received feedback from many public workers responsible for uploading payment data on behalf of their organization that they use our set of applications (e.g. aggregated profiling, error reporting) in order to identify and correct faulty data. Having stabilized the daily update process, our future direction is to interlink with foreign datasets (e.g. US, UK and Australia) to enable cross-country comparisons in the public expenditure domain.

References

1. Alvarez, J. and Labra, J. Towards a pan-european e-procurement platform to aggregate, publish and search public procurement notices powered by Linked Open Data: the MOLDEAS approach. *International Journal of Software Engineering and Knowledge Engineering,* 22, 3, 365–383, 2012.

2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., and Ives, Z. Dbpedia: A nucleus for a web of open data. In: K. Aberer, K.S. Choi, N.Noy, D. Allemang, K.I. Lee, L. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, P. Cudré-Mauroux, Eds., *Proceedings of the 6th International The Semantic Web and 2Nd Asian Conference on Asian Semantic Web Conference,* ISWC'07/ASWC'07, pages 722-735, Busan, Korea, 2007. Springer-Verlag.

3. Brickley, D. and Miller, L. FOAF Vocabulary Specification 0.98. *Namespace Document 9 August 2010 - Marco Polo Edition*, 2010. http://xmlns.com/foaf/spec/.

4. Carroll, J., Dickinson, I., Dollin, C., Reynolds, D., Seaborne, A., and Wilkinson, K. Jena: implementing the semantic web recommendations. In: *Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters,* (WWW Alt. '04), pages 74-83, New York, NY, USA, 2004. ACM.

5. Charbonneaux, J. and Kouskou-Giannakou, P. Le journalisme de “données”, une pratique d’investigation? Discours allemand et grec en regard: routines renouvelées, propos identitaire pérennisé? *Mejor, Natal (Rio Grande do Norte),* 7-10 mai 2013, 2013. http://halshs.archives-ouvertes.fr/hal-00918029/.

6. Galiotou, E., Fragkou, P. Applying linked data technologies to Greek open government data: a case study. In: G. Giannakopoulos, D.P. Sakas, D.S. Vlachos, D. Kyriaki-Manessi, Eds., *Proceedings of the 2nd International Conference on Integrated Information* (IC-ININFO 2012), pages 479 - 486, Budapest, Hungary, 2012. Elsevier.

7. Hendler, J., Shadbolt, N., Hall, W., Berners-Lee, T., and Weitzner, D. Web science: an interdisciplinary approach to understanding the web. *Communications of the ACM,* 51, 7, 60–69, 2008.

8. Isaac, A. and Summers, E. SKOS Simple Knowledge Organization System Primer, *W3C Working Group Note 18* August 2009, http://www.w3.org/TR/2009/NOTE-skos-primer-20090818/.

9. Morsey, M., Lehmann, J., Auer, S., Stadler, C., and Hellmann, S. Dbpedia and the live extraction of structured data from wikipedia. *Program: electronic library and information systems,* 46, 2,157–181, 2012.

10. Rodríguez, J.M.Á., Ordoñez de Pablos, P., Vafopoulos, M.N., and Labra, J.E. Towards a Stepwise Method for Unifying and Reconciling Corporate Names in Public Contracts Metadata: The CORFU Technique. In: E. Garoufallou, J. Greenberg, Eds., *Proceeding of the 7th Metadata and Semantics Research Conference,* MTSR 2013,

pages 315-329, Thessaloniki, Greece, 2013. Springer-Verlag.

11. Schreiber, G. and Wielinga, B. CommonKADS: A comprehensive methodology for KBS development. *IEEE Expert 9*, 6, 28–37, 1994.

12. Vafopoulos, M., Meimaris, M., Papantoniou, A., Anagnostopoulos, I., Alexiou, G., Avraam, I., Xidias, I., Vafeiadis, G., Loumos, V. Publicspending. gr: interconnecting and visualizing Greek public expenditure following Linked Open Data directives. In: *Proceedings of the W3C Workshop on USING OPEN DATA: policy modeling, citizen empowerment, data journalism,* The European Commission's Albert Borschette Conference Center, Brussels, Belgium, 2012. http://www.w3.org/2012/06/pmod/pmod2012_submission_32.pdf.

13. Vafopoulos, M., Rodríguez, J., Meimaris, M., Xidias, I., Klonaras, M., and Vafeiadis, G. Insights in Global Public Spending. In: M. Sabou, E. Blomqvist, T. Di Noia, H. Sack, T. Pellegrini, Eds., *Proceedings of the Proceedings of the 9th International Conference on Semantic Systems,* I-SEMANTICS '13, pages 135-139, Graz, Austria, 2013. ACM.

14. Vafopoulos, M. The Web economy: goods, users, models and policies. *Foundations and Trends® in Web Science* 3, 1-2, 1–136, 2011.

15. Vafopoulos, M.N., Anagnostopoulos, I., Meimaris, M., et al. ‘Storytelling’ in the Economic LOD: The Case of Publicspending.gr. In: *Proceedings of the W3C Open Data on the Web Workshop, Campus London, Shoreditch*, 2013. http://www.w3.org/2013/04/odw/odw13_submission_30.pdf.