Publishing Linked Data from existing governmental datasets

Varitimou Anastasia

SID: 3301110014

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Information and Communication Systems

OCTOBER 2012

THESSALONIKI – GREECE

Publishing Linked Data from existing governmental datasets

Varitimou Anastasia

SID: 3301110014

Supervisor: Assoc. Prof. N. Bassiliades

Supervising Committee Members: Assoc. Prof. N. Bassiliades (supervisor)

Dr V. Peristeras (committee member)

Assoc. Prof. A. Vakali (committee member)

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Information and Communication Systems

Abstract

This dissertation was written as a part of the MSc in ICT Systems at the International Hellenic University.

As more and more data is being opened up via national and local initiatives worldwide, often free of charge and without administrative overhead, its consumption in applications useful for the public, has become an index to justify the effort made.

However the amount of applications developed that use this data seems relatively low compared to the authorities' effort. Although there is an ongoing discussion on the reasons of this relatively low usage of the already available open data, here we focus on the difficulties for data integration across domains and political boundaries. This lack of integration hinders the development of high-value applications as it makes any type of data aggregation costly and challenging. The different format of the data, the lack of sufficient metadata, the different semantics and the lack of provenance information are important barriers that contribute negatively to the data integration potential.

The solution that we investigate in this dissertation is the transformation of Open Government Data to Linked Open Government Data as a means to eliminate these data interoperability and reusability barriers. Linked Data approaches allow the publication of data in semantically enriched and machine-processable formats using Semantic Web standards like RDF and SPARQL. This is coupled in our work with the use of widely-accepted and flexible vocabularies describing the content (or metadata) of the published data. In this way, the Linked Data paradigm will help overcome the problem of different data formats, and will also make possible to interlink and integrate the data with other data sources that are already available on the Web creating and contributing to what has been called as "the *Web of Data".*

This interlinked data ecosystem could then easier facilitate the development of end user applications to enable and further promote government transparency, citizen's participation, new innovative services and business opportunities.

Varitimou Anastasia

12/6/2012

Acknowledgements

I would like to express my gratitude to Dr. Vassilios Peristeras who gave me the opportunity to work with him and supervised my dissertation. Also, special thanks to Nikos Loutas, Principal Advisor at PwC, for his help, guidance and support from day one and throughout the whole thesis. They are both experts in the field and it was an honor and a pleasure to work with such a team.

I would also like to thank Agis Papantoniou and Marios Meimaris, members of the Publicspending team for cooperating with us.

But most of all, I must thank my family for the incredible patience they have shown the year of my master and especially my husband George who stood by me and supported me in every aspect in a day-to-day manner. Without you George, nothing would have been realized. Thanks a million!

Contents

List of Figures

List of Tables

1 Introduction

Access to information created or stored by public bodies is considered to be the key of success in the digital area and the lack of its availability becomes a serious problem in sound decision-making, transparency and creating new services. Since governments and authorities play major role in collecting, storing and analyzing information in a variety of fields, opening these informations to the public will launch a new era in the digital world.

Convinced of the benefits of the proactively releasing this data to the public [1, 2], governments around the globe are publishing data from various domains (financial, statistical, geospatial, legal e.tc.) under open licenses and without restrictions. It is expected that people will use it in innovative and useful applications that will increase transparency and accountability of the public sector and create new services and economic growth.

Obstacle in the optimal use of this Opened Government Data is the fact that it comes from different systems, with different identification and access mechanisms, accompanied with metadata of low quality, limited provenance information and is represented with various formats [8, 22].

The use of standards in Open Governmental Data will help overcome these fragmentations. At 2009, Tim Berners Lee invited governments not only to use open standards but also to follow Linked Data Principles in publishing governmental datasets [15]. These principles provide guides on how Semantic Web technologies can be used for the identification, accessing and retrieval of data and how to create links between them to form a giant graph called the *Web of Data.* Linked Data applications, browsers and search engines will work on top of *Web of Data* discovering data by following links. This data can be published in a modular way by "small pieces joined together" and carry semantics to facilitate large scale integration and re-use.

This thesis demonstrates how Linked Data and specifically the use of widely-accepted and reusable vocabularies, can contribute to the Open Government movement. It can

provide a standardized manner in publishing Open Government data, facilitate consensus and semantic interoperability and enable the integration of different datasets with a final goal to motivate the development of more applications that will add value to the open data initiatives.

We considered that it would be useful to explore how Open Government Data is currently being consumed in end-user applications. This research helped us to identify trends and future opportunities in the open data world but also to identify gaps and problems that application developers are facing. We realised that some of these problems could be addressed using semantic web standards like RDF and SPARQL and by following Linked Data Principles.

Since various tools for publishing and consuming Linked Data are already available, an overview of these tools together with the way that they can be used is presented.

Finally, a pipeline for publishing and consuming Open Government Data as Linked Data with the use of European Union Core Vocabularies reveals how semantic web technologies and standards can be implemented to publish *semantically rich* and *highly reusable* data, and create useful applications to the public.

1.1 Dissertation Outline

In Chapter 2 the Open Government Data movement is presented along with the most significant efforts worldwide. For the purpose of this master thesis Greek Open Government Data is used and for that reason, we give an overview of the way Greek government has embraced the Open Government movement. We also define the "Linked Data" concept and the needs that lead to the transformation of Open Government Data to Linked Government Data.

We argue that applying Linked Data techniques further facilitates the consumption of Open Government Data and in that aspect, in Chapter 3, we include an overview of software tools that have been developed to support every phase of a Linked Government Data lifecycle.

In Chapter 4, a research for identifying the way Open Government Data are currently being consumed in end-users applications worldwide is presented. The research yield more than 350 Open Data applications that were classified based on their types, categories, delivery models etc. and gave us a holistic view of the Open Government Data application world.

Chapter 5 describes a pipeline for publishing and consuming Linked Data modelled with the European Union Core Vocabularies. Key aspect in our publishing pipeline is describing our data in interoperable way based on standards to increase their reusability and for that reason we used existing vocabularies that have been developed under the Interoperability Solutions for European Public Administration (ISA) Programme from the EU, and namely the ISA Core Vocabularies. The description of the Core Business Vocabulary is also included in this Chapter.

Every phase of the pipeline is described in a separate section along with the methodology, the software tools that have been used and the outcomes of each phase. At the end of the chapter, an overview of the whole pipeline is provided together with the results and a general view of what has been achieved.

We conclude by drawing conclusions and by analysing results in Chapter 6.

2 Background and Problem Definition

2.1 Open Government Data worldwide

On his first day in the White House, the American President Barack Obama signed "The Memorandum on Transparency and Open Government" [3]. Through this memorandum the US government announced its intentions to promote openness to the public with final goals transparency, collaboration and citizen's participation. A few months later the UK government announced the launch of her Open Government Program with Tim Berners-Lee and Nigel Shadbolt advising. The Open Government movement had started to launch.

In 2011 the Open Government Partnership [4, 5] was formed, initially with the participation of eight countries (Brazil, Indonesia, Mexico, Norway, Philippines, South Africa, United Kingdom, and United States) and was steadily expanded to 55 countries nowadays. This partnership embraces countries that undertake and complete specific action plans to enter the Open Government movement.

Key factor in the Open Government movement succeeding its final goals is to make informations available to the public. With the term *Open Government Data* (OGD) we describe the various types of data that a government is opening to the public so they can be used by different groups in many different ways depending on their needs without any technological, legal, financial restrictions. It is a part of the Open Data movement but the data that are being opened are coming from government and authorities initiatives and are not published from individuals, companies or private organizations. Apparently data with personal information or national security issues are not included [5, 6, 21].

The values that Open Government Data are creating are unquestionable [1, 2, 6, 7, 8]:

- Transparency and enhancement of democratic procedures
- Participation and citizens involvement in political and social life
- Less costly and more effective government services

- New innovative services and economic growth
- Knowledge created by the availability of big volumes of data and their combination in various forms

OGD are usually provided through web portals to facilitate citizens on easier access and decrease the difficulty of searching across different types of existing informations help by numerous public sectors.

USA government launched their *data.gov* open data portal in 2009 [23], followed by the British *data.gov.uk* portal [20] in 2010. Since then the OGD movement gained momentum in many countries worldwide. From Brazil, France, Belgium, Netherlands, Spain, Estonia, Norway to Singapore, New Zeeland, and Australia enormous amount of data is being opened through government portals. Cities and municipalities like Ottawa, New York, Berlin, Singapore Chicago and Stockholm have also opened their data through web portals to facilitate their residents.

The European Union has Open Government Data also high in their *Digital Agenda for Europe* [9] and in their *eGovernment Action Plan 2011 – 2015* [10]. Through their 2003 Public Sector Information (PSI) Directive [11] and its revisions [12], they provide legislative framework in how the public sectors of the Member States should make their information available to the public for re-use, and in the 2011 Open Data Package [13] provide measures to overcome existing barriers and fragmentations across the EU. The Commission has announced a data portal in 2012 for Commission- held information and a pan-European data portal serving as a single point for EU and Member States information by 2013.

Greece was one of the first countries to embrace the Open Government movement and has introduced a set of initiatives to increase the levels of transparency, to improve public services and establish a "social contract" with its citizens [14]. Since in this thesis we are going to use Greek Open Government Data, an overview of the Greek Open Government initiatives will be presented in the next section.

2.2 Greek Open Government Movement

Greece is officially a member of the Open Government Partnership (OGP) [4] since April 2012, when its official Open Government plan [14] was accepted in the annual meeting in Brazil. Through this plan, the Greek government presented their up-to-date

efforts and their OGP commitments (Open up data, enhance public resource management and transparency and boost public engagement).

Some of the most important Greek actions towards open governance[1] can be summarized as followed:

- *Open e-deliberation and recruitment*[2] to fulfill the citizens' needs for timely information for recruitment of administration officials as well as their participation into public affairs. Citizens can comment and suggest on legislations drafts prior to their finalization.
- *The Transparency Program (Di@vgeia)*[3] that is undoubtedly, the most important initiative of the Greek government towards transparency is Di@vgeia (Di@vgeia is the Greek word for clarity). It introduces law and technical measures so that every decision from the public bodies, should first be published in the Di@vgeia platform otherwise it would not be implemented. Every decision entered in the platform is available in HTML form for anyone to see. To enable retrieval and use in applications, every decision is described with metadata and can be accessed through an API

  The API offers Rest-like calls and produces outputs in XML or JSON. Data offered by Di@vgeia's API is under Creative Commons - Attribution license, so it is available for anyone to use and modify it with the obligation to refer to the original source. Searching capabilities are enabled through a URL with the following format:

  http://opendata.diavgeia.gov.gr/api/decisions?param1=value1&param2=value2

  The primary codification system of the taxonomies is fully described in XML and various parameters are available.

Greece has also started to embrace the *Open Government Data* (OGD) movement by providing openly to the citizens:

---

[1] Greece Open Government actions http://www.ogp.opengov.gr [Accessed July 2012]

[2] Website for *Open e-deliberation and recruitment* http://www.opengov.gr [Accessed July 2012]

[3] Website for *Di@vgeia* http://diavgeia.gov.gr/ [Accessed July 2012]

- *Geodata*[4] through a central portal. *Geodata* was one out of eight governmental services worldwide providing open geospatial data to citizens and has proven extremely useful and money-saving for individuals and for the public administration.
- *Open data* regarding prices, as collected by *Prices Observatory*[5].
- *Open Taxation Data*[6] that aims to increase transparency and reduce bureaucracy in the Greek taxation system. The General Secretariat of Information Systems of Ministry of Finance has published extensive statistical data from 2000, offers in daily basis all the regional tax office's outstanding and most importantly offers Web services to other public organizations and to individuals for tax data retrieval. These services are cross checked with other Greek ministries and are not only be used as bridges with other Greek public information systems but are provided to individuals to promote transparency and facilitate private initiatives.

All the Greek government initiatives reveal their clear intention to embrace the Open Government movement and their decisive steps to create a new public administration model that will introduce significant levels of transparency, accountability and citizen engagement [14].

A prominent example of the use of Greek OGD is *Publicspending.gr. Publicspending.gr*[7]*, a* promising web project is a research initiative of the Multimedia Technology Laboratory, School of Electrical and Computer Engineering at the National Technical University of Athens. *Publicspending.gr* visualizes Greek public spending facts with central goal to promote transparency. It is in the same spirit with *Wheredoesmymoneygo.com*[8] that visualizes UK spending and the worldwide organization *Openspending.org*[9] where you can find information about government finance around the globe. Publicspending's data is being taken from Di@vgeia's API, is further validated by the Greek tax data (TAXIS) and is interconnected with foreign expenditure. Their goal is

[4] Website for *Geodata* http://www.geodata.gov.gr [Accessed July 2012]

[5] Website for *Prices Observatory* http://www.e-prices.gr/search [[Accessed September 2012]

[6] Website for *Open Taxation Data* http://www.gsis.gr [Accessed September 2012]

[7] Website http://publicspending.medialab.ntua.gr [Accessed September 2012]

[8] Website http://wheredoesmymoneygo.org/ Last accessed October 2012

[9] Website http://openspending.org/ Last accessed October 2012

enhanced by the fact that they have described Greek public spending data as Linked Data which are offered through a SPARQL endpoint.

2.3 Going from Open Government Data to Linked Government Data

The result of the OGD movement efforts worldwide is that tens of thousands of datasets are being offered mainly through portals; authorities expect people to access, analyze, integrate, use and transform them from raw datasets and machine readable formats to end-user applications with human friendly informations and facilities.

In June 2009, when the Open Government movement was launching, Sir Tim Berners Lee urged authorities not only to open their raw data to the public but also to publish them as Linked Data:

"Government data is being put online to increase accountability, contribute valuable information about the world, and to enable government, the country, and the world to function more efficiently. All of these purposes are served by putting the information on the Web as Linked Data. Start with the "low-hanging fruit". Whatever else, the raw data should be made available as soon as possible. Preferably, it should be put up as Linked Data. As a third priority, it should be linked to other sources."[15].

Nowadays billions of data has been published as RDF triples both in data.gov.uk and data.gov, SPARQL endpoints are available providing access to these triples and interesting applications have been built to demonstrate their added value.

At this point an overview of Linked Data and their principles and the possible ways they can contribute towards the goals of the Open Government movement is essential to establish the necessity of these further government efforts.

2.3.1 Linked Data

Semantic Web is the evolution of the current web in a way that informations can be interpreted by machines, so they can accomplish more complicated tasks, can access, interpret and act on these informations on behalf of the users. To achieve that, data must have a meaning for the machines and not only for people. *Linked Data* is the way to get there.

The term *Linked Data* refers to a set of techniques for publishing data on the Web, whose meaning are explicitly defined, who are related to other data sets and can be

linked to other data sets as well, following specific principles that Tim Berners Lee introduced in 2006 [16]:

1. Use URIs as names for things
2. Use HTTP URIs so that people can look up those names.
3. When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
4. Include links to other URIs, so that they can discover more things.

According to the first principal, URI's are used to identify any *thing* on the world; in the second rule, the HTTP protocol is suggested as a mechanism to retrieve these URI's, making them *dereferencable*. The third rule introduces RDF as an open standard to describe any structured data on the Web in a form of a triple:

Subject, identified by a URI, is the resource described. The object may also be described by a URI or can be a literal. The predicate defines the relationship between the subject and the object and it is also a URI pointing to a specific vocabulary. Vocabularies are RDF documents, that based on RDFS and OWL define and classify those terms. An effort is being made, to use existing terms from known vocabularies and if not possible, to relate new defined terms with existing ones. Examples of widely used vocabularies are listed below:

1. The Dublin Core Metadata Initiative (DCMI) which defines metadata like title, date, creator, subject and publisher for publications[10].
2. The Friend-of-a-Friend (FOAF) vocabulary for describing people and their relationships to others, including terms needed for the Social Web[11].
3. The GeoNames Ontology is a geographical database containing over 10 million geographical names[12].

---

[10] The Dublin Core Metadata Initiative (DCMI). Retrieved from http://dublincore.org/ documents/dcmi-terms/ [Accessed June 2012]

[11] Friend-of-a-Friend (FOAF) vocabulary. Retriever from http://www.foaf-project.org/ [Accessed June 2012]

[12] GeoNames Ontology. Retrieved from http://www.geonames.org/ ontology/ documentation.html [Accessed June 2012]

4. The Good Relations that describes terms like products, companies, prices which are used in E-commerce applications[13].
5. The vCard vocabulary that is used to describe addresses worldwide[14].

The fourth Linked Data principle refers to the interlinking of data. Linking our data with external data sources on the Web can be achieved by having triples where the subject's and the object's URI belong to different namespaces. This contributes to the forming of a giant graph called *The Web of Data*. Linked Data applications, browsers or crawlers operate on top of the Web of Data discovering additional informations about a resource.

An example of 3 triples is illustrated in Figure 1.

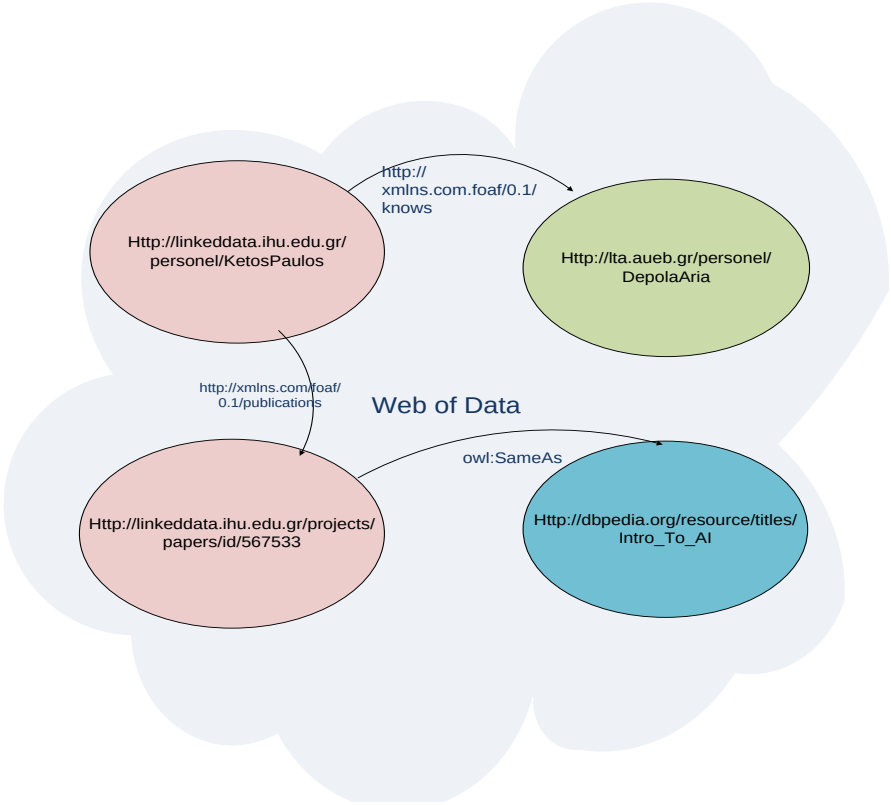


Figure 1 Example of RDF Triples

[13] Good Relations vocabulary. Retrieved from http://www.heppnetz.de/ projects/ goodrelations/ [Accessed June 2012]

[14] vCard vocabulary. Retrieved from http://www.w3.org/Submission/vcard-rdf/ [Accessed September 2012]

— The first triple states that the subject, the resource http://linkeddata.ihu.edu.gr/personel/KetosPaylos, knows the object, the resource http://lta.aueb.gr/personel/Arian_Depola.

— The second triple states that the same resource has made a publication defined by the URI http://linkeddata.ihu.edu.gr/projects/papers/id/567533.

— The third states that URI http://linkeddata.ihu.edu.gr/ projects/papers/id/567533 and the URI http://dbpedia.org/resource/titles/Intro_To_AI actually refer to the same thing, they are forming a URI alias based on the owl:sameAs property.

In our example we have 3 triples that involve 4 resources, each described by a URI; two of them belong to the same namespace. The predicates http://xmlns.com.foaf/0.1/knows and http://xmlns.com.foaf/0.1/publications are terms described in FOAF vocabulary

The evolution of the Linked Data world can best be depicted through the LOD cloud. The Linking Open Data (LOD) Project [17] kicked off on February 2007 by Chris Bizer and Richard Cyganiac with the support and sponsorship of the W3C Semantic Web Education and Outreach Group (SWEO). Main purpose was to bootstrap the Web of Data by converting datasets with open licenses to RDF triples, publish them in the Web and link them with already existing datasets. A wiki[15] is also maintained that provides information, news and statistics concerning Linked Data to the community. The LOD project has grown considerably and contained by March 2012 more than 52 billion triples from various domains (science, music, geography, government and cross-domain datasets). For a Linked Data dataset to be included in the LOD cloud, specific criteria must be satisfied (compliance with Linked Data principles, minimum amount of triples, interlinking with existing LOD datasets) and must be added to CKAN [18] , the open registry of data and content packages [19].

Figure 2[16] shows the LOD cloud with the datasets published and linked up to September 2011 colored by domain. The nodes in this diagram stands for a distinct data set published as Linked Data while the arcs represent the links that exist between the connected data sets. The figure specifies on Government linked datasets that have made a great

[15] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

[16] **Create**d based on Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. Retrieved from Website http://lod-cloud.net [Accessed June 2012]

contribution to the LOD cloud mainly due to their quantity and clear provenance [21]. In the next section we will focus mainly on this Linked Open Government Data.

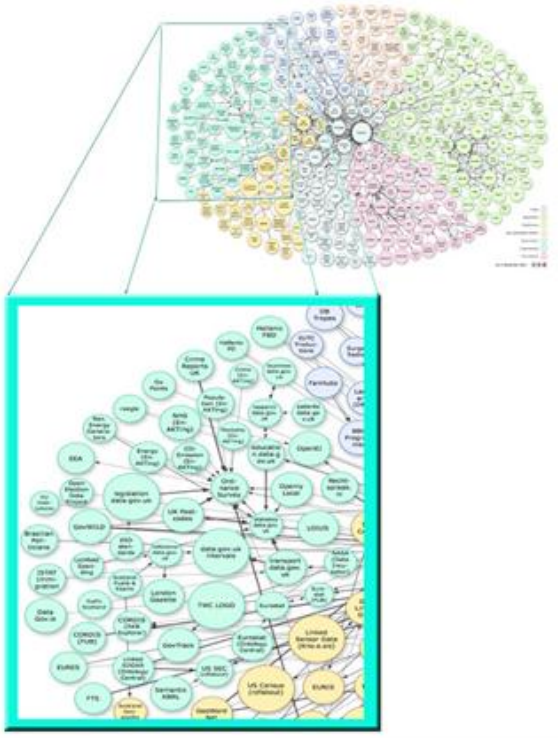


Figure 2 LOD Cloud specified on Linked Government Datasets

2.3.2 Linked Open Government Data

When Linked Data principles are applied in Open Government Data (OGD) the result is Linked Government Data (LOGD).

Tim Berners-Lee at 2010 urged Government Data owners to publish their data as Linked Data and proposed a 5-star rating system [15] described below (Table 1):

★	Make available your data on the web with open licenses
★★	Make them available as structured data (e.g. Excel instead of image)
★★★	Make structured data available as non-proprietary format (e.g. CSV instead of Excel)
★★★★	Use W3C supported open standards (RDF, SPARQL)
★★★★★	Link your Data with external related da-tasets

Table 1 The 5-star model

The first three stars are relatively easy to obtain, since most information systems outputs their data to machine readable formats and the transformation to a non-proprietary format appears to be a non-difficult procedure. For a dataset to be rated as a 4-star or 5 star, Linked data techniques should be applied.

The extra effort is essential, especially where government data is concerned [8, 21, 22]. Governments are consisting from complex organizations with multiples sectors and subsectors, usually independent from one another, and each occupies hundreds maybe thousands of people in thousands of different tasks and projects. Linked Data approach provides freedom and independency. Each organization can publish its own data, use its selected vocabularies but can match them later with new ones, if best address their needs. No holistic decision has to be taken; no holistic view has to be made upfront. The Web of Things will grow and evolve as people add *things*, links and vocabularies just like the Web of documents has done.

By September 2011, 49 Governmental datasets containing more than 13 billion triples were on the LOD cloud which represents a percentage of 42.09 % of the total LOD cloud datasets [19]. However only 4% of the triples were links to external datasets as illustrated in Figure 3.

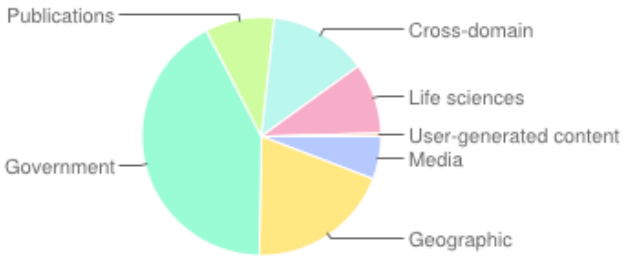


Figure 3 LOD Cloud Distribution of triples by domain, September 2011

Since Linked Data practices have already been adopted in selected governmental portals around the world, an overview of the most important initiatives would be useful.

UK's and USA's governments convinced of the benefits that Linked Data can contribute towards OGD goals adopted Linked Data standards targeting on a 4-star or even a 5-star level. They followed a different approach: in UK selected raw datasets that cover specific government domains have been transformed to 5-star Linked Data and are placed behind SPARQL endpoints or API's [20, 21, 22] while in USA they initial followed an automatic approach. The USA approach that was built with the collaboration of the *Tetherless World Constellation at Rensselaer Polytechnic Institute* resulted in the

creation of TWC LOGD portal [43] which supports the deployment, versioning and consuming of USA Linked Open Government Data. It consisted of selecting only CSV files to avoid conversions and using their own properties which were later manually linked with terms from known vocabularies [24, 25, 26]. The basic tool used for the automatic conversion and the manual enhancements, *csc2rdf4lod* is presented in Section 4.2.3.

W3C's has also done important efforts towards Linked Open Data Government movement by putting together a Government Linked Data Working Group [27] to provide standards and informations to governments that embrace semantic web technologies in publishing their data.

The current state of the above mentioned significant initiatives is depicted in Table 2.

Worldwide, governments are in the process of applying Linked Data practices to their Open Data trying to achieve a 4 or 5 star rating. Organizations, universities and researches are working on developing standards and tools to facilitate them. The efforts required are unquestionable but also are the benefits towards the Open Government movement goals.

We argue that Linked Data, especially when modelled with reusable and wide-accepted vocabularies, can empower the Open Government Data application development by providing heterogeneity in the datasets format, empower the integration of different datasets and unveil the virtues of the Semantic Web Vision.

UK data.gov.uk

UK OGD		UK LOGD		
	Number of Total datasets : 8714		**Linked data domains**	**Linked data datasets**
			Environment	Bathing Water Quality
			Finance	Combined Online Information System (COINS)
	Top Publishers		Legislation	UK Legislation
	Office for National Statistics (847)			Scottish Legislation
	Department for Communities and Local Government (739)			Welsh Legislation
	NHS Information Centre for Health and Social Care (513)			Northern Ireland Legislation
	British Geological Survey (364)		Location	Ordnance Survey Linked Data
	Centre for Ecology & Hydrology (325)		Reference	UK Government Organogram Application
	Department for Environment, Food and Rural Affairs (321)			Companies House launch of URI
	Welsh Government (241)			Guide to Companies House URI
	Department of Health (239)		Statistics	Register of Geographic Codes
	Home Office (238)		Transport	National Public Transport Gazetteer
	Department for Children, Schools and Families (227)			National Public Transport Access Nodes

US data.gov

US OGD		US LOGD	
	Number of Total datasets		**Collaboration with :**
	378,529 raw and geospatial datasets		*Tetherless World Constellation at Rensselaer Polytechnic Institute*
	Most popular departments		**Number of Linked datasets**
	Environmental Protection Agency (EPA) 1726 datasets (1594 clicks)		1.888
	Department of Veterans Affairs (VA) 308 datasets (282 clicks)		**Total Number of triples**
	Department of Commerce (DOC) 90 datasets (75 clicks)		9.951.771.397
	Department of Homeland Security (DHS) 105 datasets (62 clicks)		**Number of Linked datasets**
	Department of Justice (DOJ) 139 datasets (49 clicks)		1.888
	General Services Administration (GSA) 81 datasets (38 clicks)		**Interlinked Datasets**
	Department of Health and Human Services (HHS) 136 datasets (38 clicks)		1.686
	Department of State (STATE) 40 datasets (35 clicks)		**Links to other LOD datasets.**
	Department of the Interior (DOI) 220 datasets (35 clicks)		9.499
	Department of Defense (DOD) 33 datasets (28 clicks)		

W3C Linked Government Data Efforts

	Published Draft Standards		
	Vocabularies	Description	Websites
	Organizational ontology	Core ontology for organizational structures	http://www.w3.org/TR/vocab-org/
	Data Cube	Describing statistical data	http://www.w3.org/TR/vocab-data-cube/
	DCAT	Describing data catalogs	http://dvcs.w3.org/hg/gld/raw-file/default/dcat/index.html
	Terms for describing people	Describe people's characteristics and relations	http://www.w3.org/TR/vocab-people
	Other official working drafts:	Websites	
	Government Linked Data Best Practice	https://dvcs.w3.org/ hg/gld/ raw-file/ default/bp/index.html	
	Government Linked Cookbook	http://www.w3.org/2011/ gld/wiki/Linked_Data_Cookbook	

Table 2 Linked Data Government Initiatives worldwide [Portals accessed October 2012]

2.4 Problem Statement

Despite the large amount of Open Government Data being published worldwide there is a relatively low number of applications that uses this data to provide new services and value-added information to the users [21]. Among other reasons this is due to the fact that the multiplicity of data formats, and the absence of clear metadata and formal semantic definitions make almost impossible to combine, aggregate and mash-up these datasets as they become available from various domains and across political boundaries [8, 21, 22, 36, 53,54].

Public organizations produce large quantities of highly relevant data that is often seen as trustworthy, good quality and with clear provenance. From 2009 government and local authorities have moved decisively towards publishing huge volumes of these data. More than 700,000 datasets have been made available on the web and their usage it's not only unconstrained but strongly desirable.

However a strong backlash effect from the publishing of these until recently "kept behind closed doors" data is not apparent. Its usage in Open Government applications seems low compared with the publishing effort that is taking place worldwide.

This is mainly due to the fact that Open Government Datasets come from various organizations, are collected from different systems with various ways, contain different local and geographical characteristics and may be published in scattered catalogs. The format of this data can vary and so are the metadata that describe them. Discovering desired data and integrating it with other datasets in the same or other government portals can be challenging even discouraging.

2.5 Approach

Linked Data provides open standards for the identification, retrieval and representation of data, addresses the problem of data heterogeneity and facilitates interlinking mechanisms and discovery of further knowledge. Using wide-accepted, reusable and flexible vocabularies can further improve semantics, facilitate the discovery of joint points and integration of datasets.

Therefore, *we argue that the transformation of Open Government Data (OGD) to Linked Open Government Data (LOGD) with the use of standard vocabularies will create semantically rich and highly re-usable data. It will enable integration of datasets,*

cross-section and cross-border interoperability and facilitate the creation of new innovative applications towards the goals of the Open Government movement.

In this direction through this thesis:

1. We provide an overview of tools and technologies that have been developed for publishing and consumption of Linked Data based on the phases of the Linked Data life cycle as proposed by W3C.
2. We conduct a survey, to identify the way OGD is currently being consumed in end-user applications worldwide and to classify them according their types, the datasets they use, the delivery model e.tc. An overlook in the open data application world will help us identify the way developers have embraced the OGD movement, what they have achieved so far and how they are addressing the above mentioned problems. Trends, gaps, challenges and opportunities in the OGD apps are revealed.

   Our research was presented *at PMOD 2012 "Using Open Data: policy modeling, citizen empowerment, data journalism"* with the title "*Unraveling the mystery of Open Government Data Apps*".
3. We present a publishing and consumption pipeline which demonstrates how Linked Data tools and techniques can be applied on existing governmental datasets. Key point in our pipeline is the creation of semantically rich and high-reusable Linked Data based on EU standards by using the Core Vocabularies of the ISA Programme[17].

3 Linked Data tools

The past several years a variety of tools and approaches, have been developed to support people and organizations to better publish and consume Linked Data. In the context of the European Commission FP7 ICT Program [28], two large scale projects have been developed, to support Linked Data tools and technologies:

[17] ISA Programme Website http://ec.europa.eu/isa/index_en.htm [Accessed September 2012]

1. LOD2 [29] that aims not only to contribute interlinked versions of public Semantic Web data sets, but also to develop open source tools for data cleaning, linking and fusing and, in general, facilitating every aspect of creating new linked datasets and applications.
2. LATC (Linked Open Data Around-The-Clock) [30] is a Specific Support Action with objectives to publish European Institutions and Bodies as Linked Data, and provide a Data Publication & Consumption Tools Library, tutorials and best practices for individuals and institutions.

Especially for Linked Government Data, approaches on Linked Government Data life cycle have been developed but are not yet standardized. Figure 4 illustrates a Linked Government Data life cycle [31] proposed by Hausenblas.

Figure 4 LGD life cycle by Hausenblas, October 2011

Throughout the Linked Data life cycle, a variety of tools and techniques exist to support publishers, developers and users. For each phase different tools can be used; Table 3 depicts some of the tools and platforms which are categorized according to their functionality and use. In the following sections, we will present some of these software tools for publishing and consumption of Linked Data classified for each phase of the Linked Data life cycle.

Notice should be taken that not all the tools can be referred or described. Since the developer's community has started to embrace the Linked Data movement, dozens of tools have emerged that cannot fully be covered in this thesis. We will describe tools that are characteristic examples in each separate phase of the lifecycle.

Modeling		Create Vocabularies	Neologism, TopBraid Composer, SWOOP, MyOntology, OntoWiki, Knoodle,Protégé, MyOntology
		Re-use Vocabularies	LOV, Swoogle, Watson, vocab.org, Schemapedia
Publishing		Authoring / Validating RDF	IsaViz, Graphl, Hyena, Vapour, W3C Validator, W3C Amaya, Eyebal
	CONVERTERS	Converters for Relational Databases	D2R, Virtuoso RDF view, ODEMapster, RDBToOnto, Triplify, R2RML (mapping language), Annocultur
		Converters CSV, EXCEL	Google Refine RDF Extension, RDF123, XLWrap, NOR2O, Anzo for Excel, TopBraid Composer, CSV2RDFlod
		Converters XML, XHTML	GRDDL, Krextor, TopBraid Composer, SPARQL2XQuery, Annocultur
		Converters RDF to RDF	RDF2RDFa Converter, Mindswap, Talis Platform, AnnoCultor
		Converters to RDF for unstructured / vendor- specific data format	Virtuoso Sponger, Apenture, Anzo Unstructured, Ontos Feeder, Flickurl, Jpeg2rdf , Simile Java RDFizer, Euler-GUI
		Annotators	Open Calais , Ontos Try Workbench, Aperture, Dbpedia Spotlight, Zemanta, Faviki, One click Annotator
		Linked Data Authoring and Publishing Platforms	OntoWiki, PoolParty
	Hosting / serving	Triple stores	Virtuoso, Sesame, AllegroGraph. Mulgara, Jena TDB, Jena SDB, Garlik JXT, YARS2, BigOWLIM, 3Store and RDF Gateway
		Linked Data Interface	Puddy, Elda
		Linked Data server	Linked Media Framework
Discovery		Crawling	LDSpider, Slug
		Searching	Sindice, Swse, Falcons
		Browsing	Disco, Tabulator, Marbles, OpenLink Data Explorer
		Extracting	ANY23, Virtuoso Sponger, RDFa distiller
Integration		Vocabulary mapping	R2RFramework
		Identity resolution	Silk Link Discovery Framework, Limes
Use Case		Exploration	Sig.ma, Relfinder
		Programming libraries	Jena for Java RDFLib for Python ARC, RAP, Moriarty for PHP REDLAND,RAPTOR and RASQAL for Perl, Python, PHP, Ruby, C#, Objective-C ActiveRDF, RDF.rb for Rubby dotNetRDF, RDFSharp for .Net and C#
		Programming Frameworks	Redland RDF Application Framework for C, C-sharp, Python, Obj-C, PHP, Java, Tcl, Ruby, Perl 4Suite 4RDF, Django Web framework for Python EasyRDF, Paget for PHP Callimachus, Graphity, Sesame for Java Sparql views plugin for Drupal

Table 3 Linked Data software tools

3.1 Data awareness

The first phase of the life cycle basically concerns searching and specifying the original data that would be useful to convert to RDF and consume in Linked Data applications in the next phases. Where governmental datasets are concerned, the main government or municipalities catalogues seems as a good place to start the search for the required data. Traditional web browsers can be used; no special Linked Data software tool is required in this phase.

3.2 Modeling

In Linked data, the relationships between data are described with terms defined in vocabularies. Vocabularies are actually files written with RDFS or OWL and include all the predicates that can be used to create triples, the data types of the predicates and the relationships between one predicate and another. Identifying the right predicates to express the desired relationships and furthermore searching for existing vocabularies that contain these predicates is referred to as *modeling*. Linked data practices allow, even advices, the re-use of terms from existing vocabularies whenever the concept of the vocabulary fits the modelisation schema.

3.2.1 Re-use of existing vocabularies

A variety of search tools are available to assist the discovery of relevant vocabularies.

Linked Open Vocabulary (LOV)[18] is an ecosystem for vocabularies used in the Linked Data Cloud accompanied by metadata. List with all the vocabularies, full-text searches be based on property, element or vocabulary, SPARQL endpoint and visual navigation through links are available.

Swoogle[19] A crawler based indexing and retrieval system specialized in ontologies. Facilitates users to find an existing ontology based on terms, class or properties of a term and populates them to the user based on a ranking mechanism sorted by popularity. *Swoogle* also enables semantic web data querying with constrains on classes and properties used [32, 33].

---

[18] Website for *Linked Open Vocabulary* http://labs.mondeca.com/dataset/lov/ [Accessed July 2012]

[19] Website for Swoogle http://swoogle.umbc.edu/ [Accessed July 2012]

Vocab.org[20], *Watson*[21] and other ontology directories or semantic web search engines can be used for the exploitation of existing vocabularies.

3.2.2 Vocabulary creation tools

If no vocabulary matches the specific use case, the need of creating a new one emerges. Some vocabularies like FOAF and SIOC followed a hand-authoring process.

A simple software tool *Neologism* [22]is specially designed for Linked Data vocabularies [34]. *Neologism* is a web based vocabulary editor for creating, publishing and maintaining RDF vocabularies. It offers a limited subset of RDFS and OWL , provides URI management and HTTP content negotiations, imports vocabularies found locally or on the Web and allows the use of classes and properties from external vocabularies.

More complex ontology editors like *Protégé*[23], *TopBraid Composer*[24] and *SWOOP*[25] can also be used for the creation of a new vocabulary which in this case would be created locally and need to be applied Linked Data interfaces. Web-based systems also can be used for the deployment of vocabularies like *Knoodl*[26] and *MyOntology*[27] that provide a community based approach.

3.3 Publishing

When the vocabularies that contain the appropriate predicates have been selected, we can proceed to the next phase, the phase of publishing. Publishing refers to the conversion of the original data format to RDF and all the necessary work to be compliant with Linked Data principles. It is separated in two different sub-phases: (i) Conversion to RDF presented in 3.3.1 section (ii) Hosting the RDF and serving introduced in 3.3.2.

---

[20] Website for Vocab.org http://vocab.org/ [Accessed July 2012]

[21] Website for Watson http://kmi-web05.open.ac.uk/WatsonWUI/ [Accessed July 2012]

[22] Website for Neologism http://neologism.deri.ie/ [Accessed July 2012]

[23] Website for Protégé http://protege.stanford.edu/ [Accessed July 2012]

[24] Website for Topbraid http://www.topquadrant.com/products/TB_Composer.html [Accessed July 2012]

[25] Website for Swoop http://www.mindswap.org/2004/SWOOP/ [Accessed July 2012]

[26] Website for Knoodl http://knoodl.com [Accessed July 2012]

[27] Website for Myontology http://www.myontology.org/ [Accessed July 2012]

The software tools that can be used in each one depend on the format of the original data. The whole publishing phase for the different kind of files, from conversion to hosting is illustrated in Figure 5.

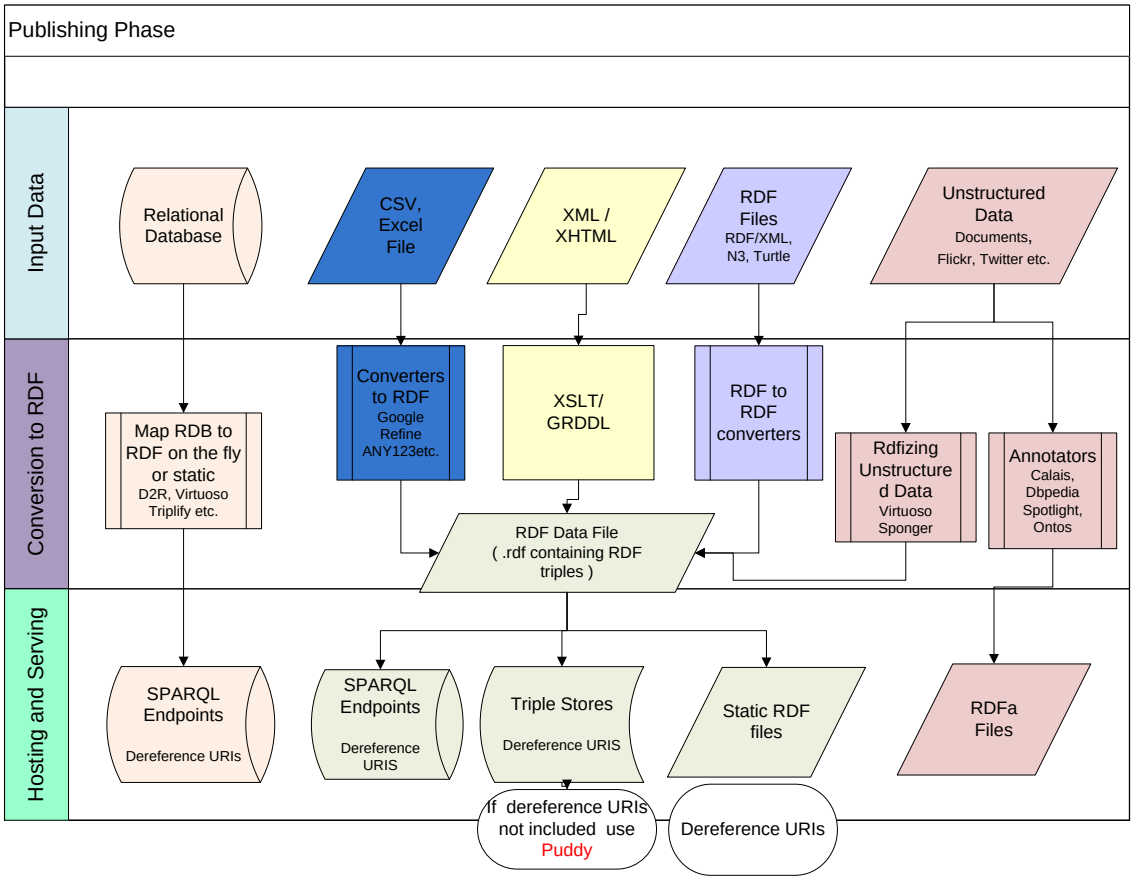


Figure 5 Publishing Phase

3.3.1 Convert to RDF

1. Publishing Data from Relational Databases

Dynamic conversion of relational data to Linked Data that is not physically generated and stored, via mapping, is widely used from most RDB2RDF systems. Important part in the transformation process of a relational database to RDF is the mapping language that will be used to convert the structure of the database to the structure of the RDF graph. Ongoing activity in W3C[28] aims to provide *R2RL* as a standard language that can be used to express customized mappings.

[28] Website for W3C R2RL http://www.w3.org/2001/sw/rdb2rdf/ [Accessed July 2012]

For publishing relational databases along with their schema as Linked Data, *D2RServer*[29]**,** which is a part of the *D2RQ* platform, is the most widely used tool. *D2R Server* uses a customized mapping created by D2RQ mapping language which enables the transformation of the original data as RDF data on the fly, without the need of keeping a dedicate RDF store. Linked Data clients and Html browsers may search and retrieve the contents of the database, while applications can use the available SPARQL endpoint to query the data.

Virtuoso's RDF view[30] service can also be used for the transformation of relational database to RDF and except the manual-written mapping language it provides, gives the possibility of automated generation of RDF links.

2. Publishing Data from structured files

Many of the raw datasets especially in government catalogues are provided in static structure formats (Excel, CSV, XML etc.). In these cases software tools, called *"converters"* or *"RDFizers"* must be used to create RDF triples based on the initial data and the appropriate terms from chosen vocabularies. A popular software tool for the RDFizing of data contained in CSV or Excel files is *Google Refine RDF Extension*[31]**,** an extension built around *Google Refine* that enables the conversion of raw datasets to interlinked RDF [35]. *Google Refine RDF Extension* supports reconciliation of imported data with related datasets on the Web spotted by search engines like Sindice, SPARQL endpoints or RDF dumps. Import of external vocabularies is available for the RDFizing process and results with their provenance can be shared in CKAN.NET.

Csv2rdf4lod[32] is another software tool that can be used to convert CSV datasets to RDF and was widely used for the RDFizing of datasets from data.gov. It provides an automatic way of converting selected datasets to raw RDF on a UNIX based system. Furthermore *csc2rdf4lod* enables manual enhancements to the subject, object, predicated of the triple, linking to external resources and in general conducts heavy-duty integration.

---

[29] Website for D2R http://d2rq.org [Accessed July 2012]

[30] Website for Virtuoso http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VOSSQL2RDF [Accessed July 2012]

[31]Website for Google Refine RDF Extension http://lab.linkeddata.deri.ie/2010/grefine-rdf-extension/

[32] Website for csv2rdf4lod https://github.com/timrdf/csv2rdf4lod-automation

A variety of other publishing tools that can be used for the transformation of CSV or Excel data to RDF also exist, like *RDF123*[33], *Anzo for Excel*[34] etc.

When the original data are in XML files, *XSLT* can be used for the necessary transformations. In these cases and in cases the data we need to convert to RDF resides in XHTM files, *GRDDL* (**G**leaning Resource Descriptions from **D**ialects of **L**anguages), a W3C recommendation, is a basic mechanism. *GRDDL* provides markups based on existing standards to ensure that the produced data are RDF compatible, for linking algorithms and extraction of data from documents. In XML files a GRDDL namespace can be included to reassure that a faithful RDF rendition of the information will be maintained in any future transformation.

TopBraidComposer's Standard Edition[35] , *Annocultur*[36] can also be used for the RDFizing of XML files or schemas.

3. Publishing Data from unstructured files

Sometimes, we want to create RDF triples from data that comes from a variety of sources like RDFa files, Web 2.0 applications like Facebook, Flickr, Delicious or may need to be acquired from specific web API's like Geonames, Freebase etc. Examples of tools that facilitate the transformation of such types of data to RDF graphs are *Virtuoso Sponger*, *Anzo unstructured*, and *Aperture*. Metadata from Web 2.0 can also be transformed to Linked Data with other specific software tools like *FlickrURL* to create triples from metadata, *Flickr wrappr* that extends DBpedia with RDF links to photos in Flickr service, *Jpg2Rdf* etc.

Virtuoso Sponger[37] is a Linked Data middleware that generates Linked Data from a big variety of non-structured formats. Its basic functionality is based on Cartridges, that each one provides data extraction from various data source and mapping capabilities to existing ontologies. The data sources can be in RDFa format, GRDDL, Microsoft Documents, and Microformats or can be specific vendor data sources like Amazon, eBay, Facebook, Digg, Delicious, Geonames and others, provided by API's.

---

[33] Website for RDF123 http://www.cambridgesemantics.comttp://rdf123.umbc.edu/

[34] Website Anzo for Excel for http://www.cambridgesemantics.com

[35] Website for *TopBraidComposer* http://www.topquadrant.com/products/TB_Composer.html

[36] Website for Annocultur http://aperture.sourceforge.net/

[37] Website for Virtuoso Sponger http://virtuoso.openlinksw.com/dataspace/dav/wiki/Main/VirtSponger

4. Converters RDF to RDF

The RDF model is based on triples (subject – predicate – object). In order to publish them on the Web, various syntaxes can be used. RDF/XML syntax seems to be the most common and it is supported by W3C. An example of 2 triples stating that someone is a person and has made a publication can be expressed in RDF/XML syntax as follows:

```
<?xml version="1.0" encoding="UTF-8"?>
 <rdf:RDF
 xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
 xmlns:foaf="http://xmlns.com/foaf/0.1/">
  <rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/personnel/KetosPaylos">
 <rdf:type rdf: xmlns.com/foaf/0.1/Person"/>
 <foaf:publications> URI http://linkeddata.ihu.edu.gr/projects/papers/id/567533</foaf:publications>
 </rdf:Description>
</rdf:RDF>
```

This syntax is not particular human friendly and sometimes other syntaxes like RDF/N3 (N- Triples) and Turtle can be used. The above triples can be expressed in RDF/N3:

```
< http://linkeddata.ihu.edu.gr/personel/KetosPaylos> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <rdf: type rdf: xmlns.com/foaf/0.1/Person"/>
< http://linkeddata.ihu.edu.gr/personel/KetosPaylos> < rdf: xmlns.com/foaf/0.1/Publications>
<http://linkeddata.ihu.edu.gr/projects/papers/id/567533>
```

Sometimes there is an apparent need for converting one RDF syntax into another. There is a number of RDF converters like for example the java tool *RDF2RDF*[38] and the on-line Web tool *Mindswap*[39].

5. Annotators

When the input data reside in plain text, software tools like *Dbpedia Spotlight*[40], *Open Calais* [41], *Ontos*[42], *One click Annotator*[43], *Zemanta* for annotating blog posts, and *Faviki* provide annotations to make the content discoverable from faceted browsers or applica-

[38] Website for RDF2RDF http://www.l3s.de/~minack/rdf2rdf/

[39] Website for Mindswap http://www.mindswap.org/2002/rdfconvert/

[40] Website for Dbpedia Spotlight https://github.com/dbpedia-spotlight/dbpedia-spotlight

[41] Website for Open Calais http://www.opencalais.com/

[42] Website for Ontos http://www.ontos.com/

[43] Website for One Click Annotator http://www.loomp.org/

tions [36]. These tools connect Web of Documents and Web 2.0 with Web of Data, by selecting desired text and vocabularies terms and incorporate semantic functionality offering a richer user experience [37].

6. Linked Data Platforms

Advanced platforms for integrating knowledge, authoring and editing RDF content from scratch, serving and consuming it in Linked data applications, have also been developed.

OntoWiki[44] which is a lod2 supported open source project facilitates visual representation of knowledge and enables RDF authoring. It can act as Linked Data Server and through his Pingback service allows the Linked Data owner to be notified about the usage of his resource.

PoolParty[45] is a suite comprising of 3 different components. The basic component manages metadata based on the Semantic Standard SKOS, integrate and publish Linked Data. The second component provides extracting capabilities and high performance data mining while the third enables semantic search.

3.3.2 Hosting and Serving

When the RDF files are ready through the conversion process, serving it on the Web is the next step. When its size is relatively small, it can be uploaded to any web server. For bigger files, storing them as RDF dumps so they can be downloaded or storing them in RDF triple stores and make them available via SPARQL endpoints, are the most common procedures.

RDF triple stores are tools that are used as RDF databases. They provide mechanisms for storing and accessing RDF graphs. A number of triple stores are currently available (see Table 2 → Triple Stores). Choosing the right one depends on the amount of RDF data and the implementations they are offering and seems to be classified in 3 major categories [38]:

- In-memory where the RDF graph is stored as triples in main memory. For example an RDF graph on Jena's API

---

[44] Website for OntoWiki http://ontowiki.net/Projects/OntoWiki

[45] Website for PoolParty http://poolparty.biz/

[All the above accessed July 2012]

- Native triple stores that have been created from scratch to act as RDF databases. This seems to be the most popular and efficient category. Virtuoso, Allegro Graph are prominent examples in this category
- Non- native triples stores that act as intermediaries on top of already existing database engines. Known example of this category is Jena SDB.

Triple stores are storing RDF data and preferable provide access to them via SPARQL 1.1 endpoints. Some triple stores provide user authentication or permissions on the level of dataset. Preferable they provide manageable interface and faceted browsing.

Well known triples stores/ frameworks are:

Jena[46]: A framework that provides storing, querying, reasoning and SPARQL endpoint. It has extensions to run as a standalone server or within another server like Apache. It contains two plug-in, *Jena SDB* that runs as a non native triple store on a conventional database and *Jena TDB* that stores and queries RDF datasets using a purpose built-in triple store

Open Link Virtuoso Universal Server[47]: a hybrid architecture that can run as storage for multiple data models, such as relational data, RDF, XML, and text documents. Virtuoso supports repository management interface and faceted browsing of the data. It can run as Web Document server, Linked Data server and Web Application server. It supports HTTP based web services including support for Representation State Transfer (REST) and Simple Object Access Protocol (SOAP).

Sesame[48]**:** Sesame supports storage, querying and inference of RDF data that can be used in any Java application. Through Sesame both in-memory and native RDF repositories can be created, query services are available and an HTTP server for accessing sesame repositories via HTTP. Sesame is open source and provides several extensions.

Choosing the right store depends on a number of criteria:

- Scalability: How many triples can it store and how many triples do you expect to have in the future.

[46] Website for Jena http://jena.apache.org/ [Accessed July 2012]

[47] Website for OpenLink Virtuoso http://virtuoso.openlinksw.com/ [Accessed July 2012]

[48] Website for Sesame http://www.openrdf.org/ [Accessed July 2012]

- Performance: *Lehigh University Benchmark* (LUBM) is a widely accepted methodology for evaluating triple-stores. The metrics in the LUBM evaluation methodology include load time, repository size, and the response time of 14 different queries [38]. The *Berlin SPARQL Benchmark* (BSBM) [39] is another benchmark for comparing the performance of Triple stores that expose SPARQL endpoints.
- Open source or Commercial: BigOWLIM and AllegroGraph are commercially available. For open source stores, note should be taken not to pay after a certain threshold.
- Support and community: Some open source may have retired so a community support will be absent
- Implementation and facilities that each triple store offers

In some cases Linked Data interfaces are not supported. In this case, *Puddy*[49] is a popular tool designed to turn a SPARQL endpoint to a Linked Data server by providing dereferencable URIs (URIs used in most SPARQL datasets are **not dereferenceable)** and handle HTTP negotiations.

3.4 Discovery

One of the foundations of Linked Data is the discovery of related URIs through links. Finding related datasets and understating the best joint points between them will enable integration, aggregation, data visualizations e.tc.

To facilitate this process several tools can be used that belong in four major categories: *Linked Data Crawlers, Linked Data Search Engines, Linked Data Browsers, and Linked Data Extractors* [30]. In the following sections we will give an overview of popular tools for each category.

Linked Data Crawlers

Linked data crawlers can be used to traverse the Web of Data by following RDF links related to "seed" URIs. The related links found can be stored in files or in RDF stores.

[49] Website for Puddy http://www4.wiwiss.fu-berlin.de/pubby/ [Accessed July 2012]

LDSpider[50] is a popular open source crawler that handles links found in different formats. It can traverse RDF/XML, N-Triple files and with communication with Any23 server can extract RDF links from RDFa and Microformat documents. The output links found can be stored to RDF/XML, N-Quads format files and alternatively they can be stored in a Triple store. For a simple crawling task *LDSpider* can be used from command line but can be used inside an application as a simple parser through its Java API.

Linked Data Search Engines

Linked data search engines crawl the Web of data by following RDF links, aggregate the resulted data and provide querying capabilities over it in a similar way to the traditional databases. For example *SWSE*[51] presents users a ranked list of entities based in specific keyword queries [40]. The description of entities is aggregated from different sources and may include inferred data through reasoning [21].

Other search engines like *Sindice*[52] also include APIs so that the querying results can be provided not only to humans but also to applications. *Sindice* is a semantic web platform that offers searching in the Web of Data based on keywords and URIs and querying based on a SPARQL endpoint that currently offers more than 12 billion triples and it is updated live. Search API that provide programmatic access to its search capabilities, cache API for read-only access to data store and Live API that allows retrieving triples from web documents, are available.

Linked Data Browsers

Linked Data browsers have been developed that traverse the Web of Data by following RDF links. *Disco*[53] is considered the navigation paradigm of Web of Data [37]. It retrieves informations about a specific resource by dereferencing HTTP URI and follows rdfs:seeAlso links. *Marbles*[54], a server-side application for XHTML clients and Tabulator[55] are other examples of linked data browsers that provide also provenance informations.

---

[50] Website for LDSpider http://code.google.com/p/ldspider/ [Accessed July 2012]

[51] Website for SWSE http://swse.deri.org/ [Accessed July 2012]

[52] Website for Sindice http://sindice.com/ [Accessed July 2012]

[53] Website for Disco http://www4.wiwiss.fu-berlin.de/bizer/ng4j/disco/ [Accessed July 2012]

[54] Website for Marbles http://marbles.sourceforge.net/ [Accessed July 2012]

[55] Website for Tabulator http://www.w3.org/2005/ajar/tab [Accessed July 2012]

Linked Data Extractors

In case our data are contained in Web documents with various formats, software tools called "extractors" are used to extract and convert it to RDF. *ANY23*[56] is a commonly used tool that parses Microformats, RDFa and Microdata and also RDF/XML, Turtle, N-Triples and Quads to converts them to RDF in various syntaxes. It is used widely in Sindice search engine and Sig.ma and can be provided as an on-line service, or as a library in Java applications.

3.5 Integration

With this phase, we increase the value of the published Linked Data by mapping the terms that were used for publishing the triples, with term in existing vocabularies (described in section *vocabulary mapping*) or by interlinking the newly published data with other datasets (described in section *identity resolution*). This process that is referred to as *integration* is the main idea behind the *Web of Data* and leads to the discovery of new knowledge and their combinations in unforeseen ways [37].

3.5.1 Vocabulary mapping

During the phase of modeling a new dataset, searching of terms from widely used vocabularies to express the relationship may not proved to be satisfactory. In these cases, new terms should be defined in proprietary vocabularies. Since Linked Data application developers expect an integrated view of their resources, mapping between these new terms with existing terms using OWL or RDFS link types, is in need. *R2R Framework*[57] is a software tool that can be used to create mapping between terms and also includes a Java API to transform the data according to this mapping [41].

3.5.2 Identity resolution

For a Linked dataset to be rated as 5-star, linking to external datasets is required. The process of discovering related datasets as linking targets, identifying related terms between these external datasets and the newly published dataset and interlinking them can be depicted in Figure 6.

---

[56] Website for Any23 http://any23.org/ [Accessed July 2012]

[57] Website for R2RFramework http://www4.wiwiss.fu-berlin.de/bizer/r2r/ [Accessed July 2012]

The interlinking process with related datasets can be performed manually or automatically. A tool to automatically interlink datasets is *Silk Link Discovery Framework*.[58] *Silk* uses *Silk-LSL* language to define the RDF types that should be discovered and the conditions they must fulfill and accesses the data sources via a SPARQL endpoint [42]. The latest versions allow the data sources to be described in RDF files. Another tool that can be used for link discovery is LIMES that utilizes metric spaces to compute estimates of the similarity between instances. It can be used as a Web application and as a standalone tool.

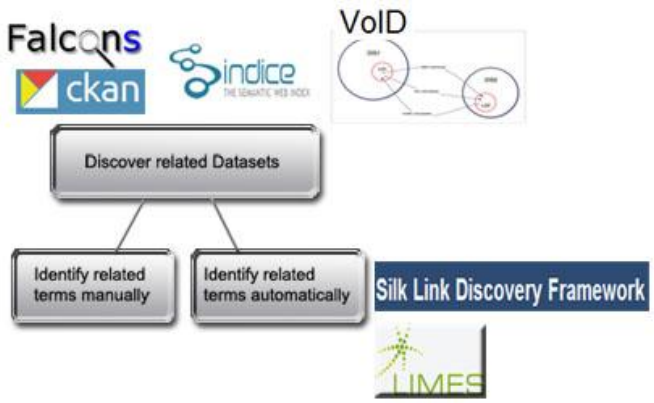


Figure 6 Discovery of related datasets

3.6 Use Cases

In the last phase of the life cycle, the implementation of a use case based on the 5-star linked datasets that have been published and interlinked through the previous phases, validates and justifies the effort made.

There are many successful use cases worldwide and a variety of tools and platforms to support their development. W3C[59] maintains a list of some of these use cases. Successful applications that use Linked Data can also be found in government and organization catalogues. Especially in the U.S data.gov [23] and TWC LOGD project [43], there is an apparent trend to stimulate and facilitate developers to use the published Linked Data. Use case demos illustrate the practical values of linked data; their technical details are offered to developers together with metadata for the software and datasets used. Tu-

[58] Website for Silk http://www4.wiwiss.fu-berlin.de/bizer/silk/ [Accessed July 2012]

[59] W3C semantic web use cases http://www.w3.org/2001/sw/sweo/public/UseCases/

torials and “Mashathons” aim to give incentives to developers to merge published linked datasets “in a way that is not expected”.

Web applications enable the exploration of the Linked Data world, like *Relfinder*[60] that provides in a visualize way the links between resources and *Sig.ma*[61] that queries Sindice to find all data relevant with a keyword or resource and aggregates them into a data list.

As for the development of new Linked Data applications, a variety of frameworks and libraries like Jena, Sesame for Java, ARC2 and Moriarty for PHP etc. give developers the ability to retrieve data from SPARQL endpoints and use them in their applications. Plugins like *The SPARQL Views plugin for Drupal*[62] enables developers of Drupal Web sites to integrate data from SPARQL endpoints.

Google Visualization API seems to be a common software tool in Linked Data applications, as visualizations are apparent in most use case. The API facilitates users to embed visualizations in their applications using JavaScript or HTML and keep them always up to-date by connecting them live to a data source.

3.7 Conclusions on Linked Data tools

Our overview in Linked Data software tools confirmed the fact that the developer’s society, having realized the great potential of publishing huge amount of data as Linked Data, has created a variety of tools to convert them to RDF quickly and efficiently. This will allow even more Linked Data to be published and create the network effect that is necessary to create mashups. Furthermore the development of ready-to-use programming libraries and frameworks, try to ease the way for building client applications mainly interfering with real-world users.

---

[60] Website for Relfinder http://relfinder.dbpedia.org/relfinder.html [Accessed July 2012]

[61] Website for Sig.ma http://sig.ma [Accessed July 2012]

[62] Website that can be downloaded https://github.com/linclark/sparql_views

4 Open Government Data applications

The level and the way OGD is being consumed will give us an initial hint of how successful the Open Government movement has been in achieving its final goals: transparency, participation and collaboration, innovative new services and economic growth. It would also help us identify gaps and challenges that could be addressed by using Linked Data technologies modeled with standard vocabularies.

4.1 Discovery of OGD apps worldwide

In the Digital Agenda for Europe, the European Commission (EC) [9] is emphasizing the importance of opening up data resources for reuse. The great potential of open public sector information is widely recognized and the impact this would have on transparency, innovation and the real economy is indisputable [1, 2].

Towards this direction, the EC encourage and facilitate Member States to open up government data and metadata [10, 12]. Governments in turn have started taking vital steps towards opening up their data. As the volume of data opened up all over the world through national or local initiatives as well as by organizations such as Eurostat and the World Bank is steadily growing, open government data (OGD) consumption and usage in different types of applications has become an index to justify this enormous effort.

In this vein, OGD providers often invite and encourage software developers to create applications that use OGD (e.g. through hackathlons and contests) and integrate into value-added applications. As a result, nowadays hundreds of OGD apps exist. However, a holistic view of the growing space of OGD apps is currently missing

In this direction, we considered that it would be useful to gather opened data applications worldwide and classify them in various ways, trying to unveil the way they were addressing problems caused by the different formats and characteristics of datasets, but also to identify trends, opportunities and if possible drawn useful conclusions. In the following sections we will present (section 4.2) the methodology that we followed, the finding from our survey (section 4.3), conclusions and future research directions (4.4).

4.2 Methodology

The survey comprises of three phases performed following a spiral approach: *(i)* Web search for available OGD apps; *(ii)* Documentation of the OGD apps found online; and *(iii)* Analysis of the collected data.

The Web search started from the app catalogues and directories of national and local OGD portals maintained by OGD champions, such as data.gov, data.gov.uk and Ottawa city. The search was expanded to cover also the journalist world that has also joined the open data initiative, e.g. the Guardian Open Platform. The open data portals of large non-governmental organizations that publish OGD, e.g. World Bank and Eurostat, were also examined. Finally, a general purpose Web search using Google Search Engine was also carried. The most frequently used search terms include: OGD, open government data, open data, app, application, linked government data, as well as their combinations.

The metadata used for describing OGD apps (usually the ones published in catalogues) were also documented. The analysis of the metadata gave us a superset of common terms that we then use for documenting the apps. Hence, the following information is kept (wherever possible due to availability constraints) for each of the OGD apps discovered: Name, URL, Description, Publisher, Catalogue, Programming Language & Framework, Reuse & Pricing (e.g. open source, require fee), Delivery mode (e.g. mobile, web, desktop), Category, Datasets used, Vocabularies used (e.g. VoID [51] or dcat [52]), Use Linked Data, Usage info (e.g. downloads or likes), Date of creation, Date of last modification, Intended audience, and Status.

Finally, the data collected is analysed in order to extract different types of conclusions and patterns about open government apps. Since a vast majority of OGD apps, if not all of them, can be accessed through catalogues, a description of these catalogues and model elements are available.

4.3 Data Analysis

The initial search of this survey yielded more than 350 apps. The majority of them were found in government and city catalogues. 46 of them came from the Guardian Open platform, 10 from the World Bank and 17 apps were found in independent sites.

We observe that by now numerous cities, including Ottawa, Toronto, New York, Seattle, Colorado, Dublin, London, Rennes, Berlin, Stockholm and Singapore, have carried out an OGD apps contest of some type to bootstrap the use of their data. This paradigm has also been followed by other types of OGD providers, such as the World Bank.

We have identified the 13 categories of OGD apps, namely health & safety, entertainment, sports, transportation, city services, real estate, environment, education, public safety & law enforcement, food & dining, development, business & finance, government & civics. This classification was created by integrating and generalizing classifications of apps found in different catalogues. Apparently, different classifications are also possible, namely according to other metadata elements used for documenting the apps.

Figure 7 illustrates the number of OGD apps per category reviewed so far. We observed that local initiatives such as hackathlons and OGD app challenges organized by cities delivered apps mostly related to entertainment and transportation. This justifies the fact that these two categories, followed by environment and government & civics, are the most popular ones.

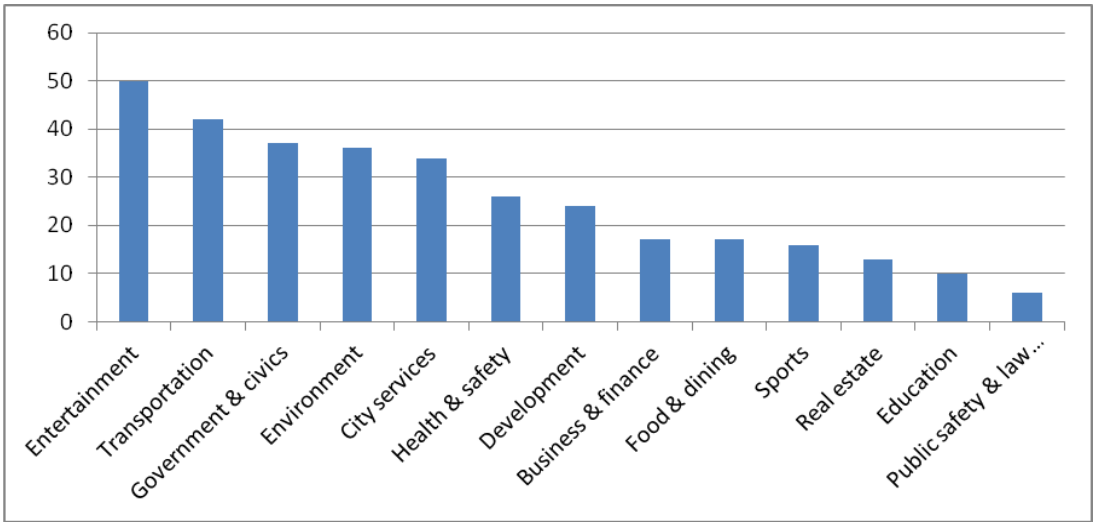


Figure 7: Number of OGD apps per category

The vast majority of OGD apps have been developed by individual developers, freelancers and research institutes. This finding indicates that the business community is not yet convinced about the business and economic potential of OGD, and are thus reluctant to experiment with apps which could produce revenues and increase their firm's visibility in a rapidly evolving community.

Currently over 90% of OGD apps are available for free. Few of them (found in App Store and in Android Market) require a monthly subscription fee.

Most of the apps are Web-based followed by apps developed on proprietary mobile frameworks, mainly iOS, Android and Windows Mobile. Some apps can be deployed on different platforms.

OGD providers develop usually support apps and services (e.g. publish/subscribe mechanisms and rss feeds), focused mainly on facilitating the access to OGD and on communicating changes (e.g. release/update/deprecation of a dataset).

OGD apps are all about data visualization. Different ways of visualizing data are come across, e.g. using graphs, maps, tables etc. For example, apps classified under health & safety use graphs and charts to visualize information related to obesity statistics, smoking and non-smoking venues, infant mortalities and the spread of diseases. Along the same lines, Government & civics apps use visual methods to bridge the gap between citizens and central government and deliver PSI in an easy to digest manner. For example, they visualize the votes of elected representatives on specific bills, and the spending of both public bodies and individual elected representatives.

Approximately 90% of the apps surveyed combine OGD with maps, mainly Google Maps and Open Street Map, in order to visualize data and provide value-added location-based services. Many of these apps are also context-aware, i.e. they identify the exact location of the user either by using GPS coordinates or by searching based on postcodes, thus offering a richer and personalized user experience. For example, a significant part of the apps under health & safety locate hospitals, health facilities, emergency facilities (e.g. firehouses), pharmacists and care homes and guide to the user to the closest one. In order to select the closest facility they either use GPS coordinates for finding the user's location or support search based on postcodes. Likewise, entertainment includes a wide variety of apps, e.g. for finding parks, special events, theatres, gyms, points of interest, shops and all sorts of recreational activities, while city services location-aware apps allow citizens to find the closest recycling point or provide local employment market information.

Most of the OGD apps rely on static datasets. However, few of them consume real-time OGD. These are found under transportation, environment and public safety & law enforcement. For example, many apps under transportation visualize traffic information on city maps or inform the user on expected arrival times of busses and trains. Other apps under environment consume real-time UV and pollution data to warn the user in cases of emergency, while apps classified under public safety & law enforcement use real-time OGD to inform the user on crimes, accidents or other types of emergencies that have occurred within a specific distance and/or timeframe.

We observe that a significant portion of OGD apps rely on a single dataset, while few of them integrate more than one datasets. In our view though, the power and the real value

of OGD can only be released through the integration of complementary datasets. We expect OGD integration in apps to gain popularity as linked data and semantic technologies become more mature. Currently, however, only few apps use these technologies.

Some apps integrate OGD with data coming from the Social Web, e.g. user's opinions or even data from Wikipedia. For example, real estate apps combine OGD with user's preferences or engage citizens in discussions around urban development. Likewise, food & dining apps combine OGD with user's ratings in order to rank the quality and the cleanliness of restaurants.

Finally, a number of apps classified under city services and government & civics use OGD and visualization technologies in order to facilitate the communication between (local and/or central) government and citizens and to engage citizens in politics. For example, citizens can use OGD apps to report to the city issues related to graffiti, potholes, excessive garbage, street problems or street lighting. Other apps enable citizen engagement and participation by facilitating citizens to express opinions, interact with elected representatives, raise issues for discussion and comment on the activities of the (local and/or central) authorities.

4.4 Conclusions

The outlook of OGD apps is positive; however certain factors that will impact their future growth are identified.

Currently, finding an OGD app that will fulfill the needs of an individual is not a trivial task. It may be published in an OGD app catalogue or it may be available anywhere on the Web just as well. How would a common metadata model for describing OGD apps and lightweight semantics facilitate OGD app discovery?

OGD apps can be made available following different exploitations roots (from commercial applications to free and/or open-source ones. But can traditional business model fulfil the peculiarities of OGD apps?

By exploring existing apps in detail, one can collect valuable information related to the most used datasets, the most popular types of apps and the lifecycle of an app.

Form our research we can conclude that integration of datasets is not the prevalent trend in OGD apps. However, we argue that data integration will unveil the real value of OGD. Additionally OGD is often made available in non-machine-readable formats, which also hampers ODG use and integration.

Linked data with the use of standard vocabularies have the potential to overcome the problem of data integration. In the next chapter a pipeline will demonstrate how Linked Data principles with the use of wide-accepted vocabularies can create highly re-usable data, enable data integration and lead to the development of more application towards OGD movement goals.

5 Pipeline for publishing and using Linked Government Data

In this section we will describe a pipeline for publishing *Greek companies and public bodies* as highly reusable Linked Data using the proposed by EU *Core Vocabularies* and through a faceted browser we would make them available to end-users.

A description of the Core Vocabularies and specifically the Business Core Vocabulary will be available in section 5.1.

In section 5.2 we will describe our case study and the initial problems that led us to describe all Greek legal entities in a way that will obtain high reusability and cross-border interoperability.

In section 5.3, each phase of the publishing and consumption pipeline will be presented. The pipeline will be based on the Hausenblas Linked Data lifecycle and the software tools used for each phase have been identified as more suitable through our research described in the previous section. Finally, the results of our pipeline will be presented in section 5.4.

5.1 Core Business Vocabulary

European companies are highly conducting business across national borders and equally significant is the need of communication in the public administrations in the Member States. Obtaining informations of business partners in a standardized way is in need and the *Core Business Vocabulary* can be the basis for it.

Since in this thesis, Linked Data principles are going to be applied to data taken from the Greek government, modeling them with terms from EU vocabularies will provide our dataset with cross-domain and cross-border semantics. In this vein, we decided to use the Core Vocabularies proposed by the EU as basis for our schema. A brief description of the three vocabularies and a more detailed one for the Core Business vocabulary will follow.

The Core Vocabularies are being introduced through Action 1.1[63] of the European Commission's ISA (Interoperability Solutions for European Public Administrations) program[64] that promotes semantic interoperability in the European Union Member States. Three e-Government *Core Vocabularies (Person, Business and Location)* have been developed to support public sector information exchange between the 27 Member States. They are addressed as *Core* because they describe the minimal, basic characteristics of the entity; in this way they can be highly reusable and extensible. Compliance with the Core Vocabularies provides a common starting point and guarantees interoperability, while specializations can be built by adding metadata to the core [44].

Figure 8 depict the conceptual model of the three *Core Vocabularies*, which as we can observe have strong links between them. *The Core Person Vocabulary* [46] describes people as a natural entity, *Core Location* [47] describes places, addresses and geographical coordinates and *Core Business Vocabulary* [48] aims to describe businesses as a legal entity; organizations, their hierarchies and traders that do not have a single legal entity cannot be described.

[63] Website for Action 1.1 http://ec.europa.eu/isa/actions/01-trusted-information-exchange/1-1action_en.htm [Accessed July 2012]

[64] Website for ISA program http://ec.europa.eu/isa/ [Accessed July 2012]

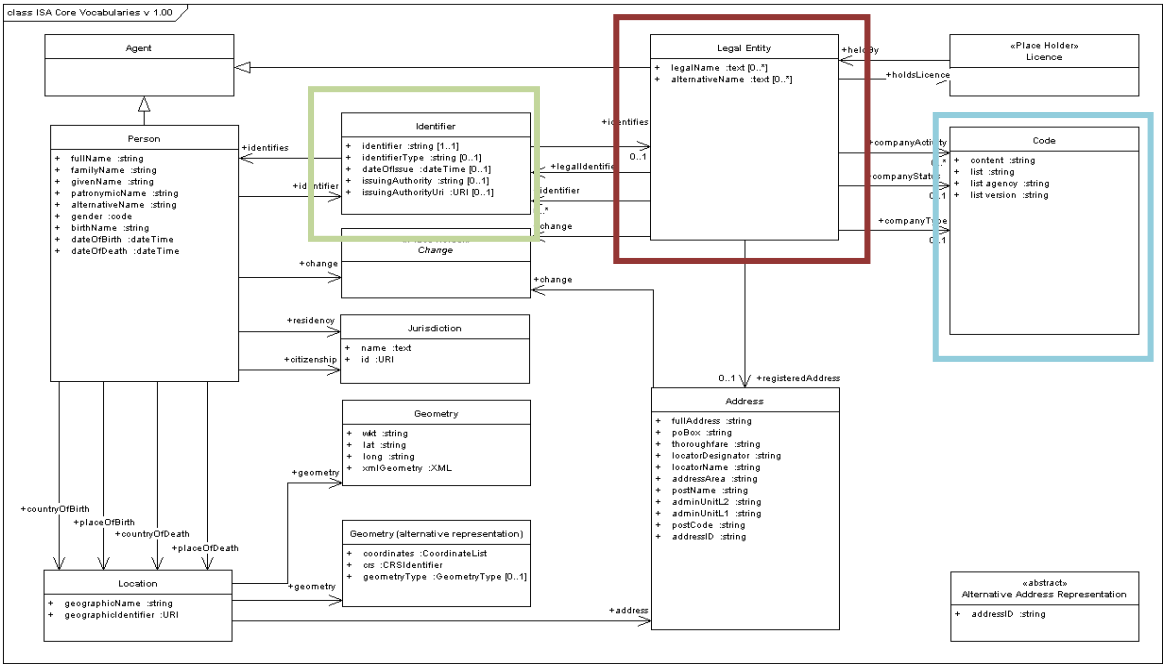


Figure 8 Core Vocabularies UML Diagram [45]

In Core Business Vocabulary the key class is Legal Entity and represents a business. The properties each Legal Entity may have are described in Table 4.

legal name	The legal name of the business as it is registered in nation business registries
alternative name	If there is any but cannot be used for translation reasons
company type	The representation of company types differs across countries. The limited set of company types of each country should be used (In Greece: ΕΠΕ, ΑΕ, ΟΤΑ, ΝΠΔΔ etc.)
company status	There are no specific terms to describe the status of a company cross-border. "Normal Activity" is proposed. "Bankrupt", "Inactive" are other possibilities.
Company activity	The use of NACE codes is suggested to obtain EU interoperability if no other national system exist
Legal Identifier [1..1]	Each company **must** have **one** legal identifier which the basic relationship between a legal entity and the authority with which it is registered. In Greece *AFM* should be used as it is a unique representation of a legal entity.
Identifier	Other identifiers except from the previous legal identifier may also exist
Registered Address	This should be the public, registered address of the business. This is actually a sub property of the address. The actual address should be described with the Core Location Vocabulary

Table 4 Core Business Vocabulary – Legal Entity Properties

In Core Business Vocabulary there also two other classes included:

The Identifier class captures the concept of the legal and other identifiers. Its properties are described in Table 5:

Identifier	skos:notation to provide the actual identifier
Identifier type	dcterms:type to provide an identifier for the type of identifier issued
Date of issue	use dcterms:created to provide the date on which the identifier was issued
Issuing authority	adms:schemeAgency to provide the name of the agency that created the identifier (as an rdfs:Literal)
Issuing authority URI	dcterms:creator to link to a URI of an Agent class describing the issuing Agency

Table 5 Core Business Vocabulary – Identifier Properties

The *Code* class can be used to describe properties like *company activity*, *type* and *status* that are differently represented across borders. If we used a literal "ΝΠΔΔ" to describe the *company type* of a public entity in Greece, it would have no meaning in the other countries since public entities are represented in a different way. To obtain interoperability it would be preferable to use terms from controlled vocabularies instead of literals. If such a vocabulary does not exit, it should be created as part of the dataset using the SCOS concept in a way that it will be compatible with the Code Class, as it is depicted in Figure 9.

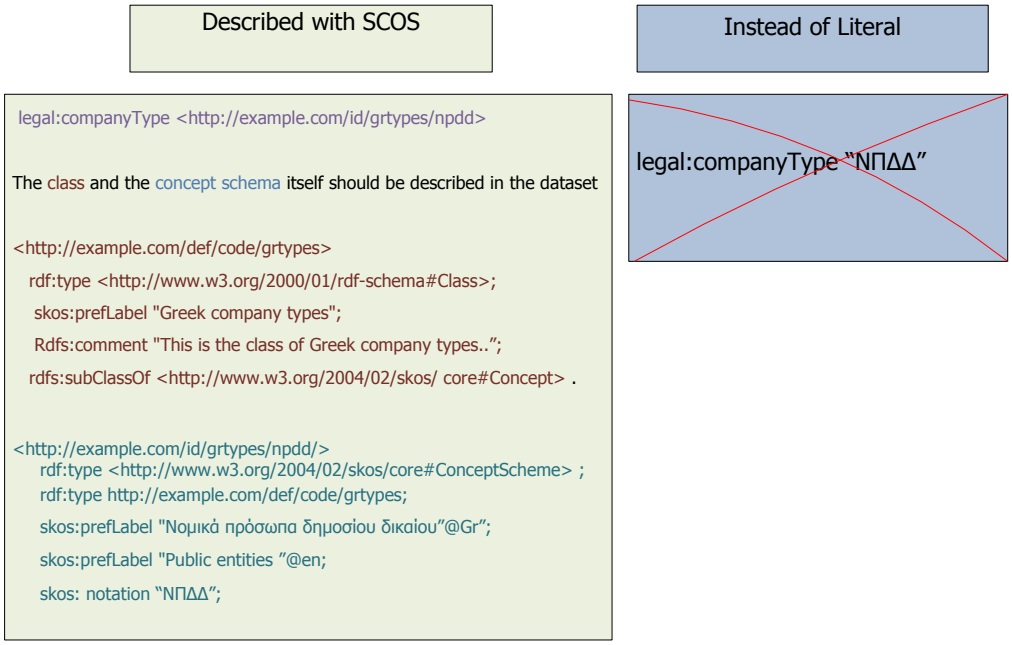


Figure 9 Defining Class and Concept schema for Company Types

Central point in the Core Business Vocabulary is that since every legal entity is conferred by an authority, it must have one ([1..1]) legal identifier which can vary depending on the country's business register. In Greece the tax number (AFM), which is unique for every legal entity, plays the role of the legal identifier.

The URI's for the three vocabularies can be seen in Table 6 and each class and property URI can be constructed by appending the term after the hash (e.g. http://www.w3.org/ns/legal#LegalEntity)

Core Person Vocabulary	http://www.w3.org/ns/person#
Core Business Vocabulary	http://www.w3.org/ns/legal#
Core Location Vocabulary	http://www.w3.org/ns/locn#

Table 6 URIs for the Core Vocabularies

5.2 Case study

Greek legal entities (companies, public bodies etc.) are described in various catalogues and applications, each one using a different format and accompanying them with different metadata. For instance:

(1) The Greek legal entities' formal source is Taxis.net from the General Secretariat of Information Systems of Ministry of Finance which although they provide openly this data to citizens, they describe it by following their own proprietary data model.

(2) Di@vgeia that by law contains all decisions about every Greek public body and therefore contains informations for almost all legal entities involved, is describing them with its own model.

(3) Publicspending.gr that is built on Di@vgeia and on Taxis' data uses a slightly different model.

As result:

A. Neither (1), (2), (3) are standardised, they are hence non-interoperable with each other and with other business data formalisms.

B. There exists no reference hub that is a service trusted LOGD about Greek companies.

We address these problems by transforming Greek legal entities to Linked Data modelled with terms taken from standard vocabularies and more specifically with the proposed by the EU Core Business and Location Vocabularies. Core Vocabularies have specially been designed to build consensus on diverse metadata and reference data not only between different organizations and domains but also between the different Member States.

The whole process of transformation will be described through a publishing and consumption pipeline. To reveal the true benefits of our data, we would identify and create links with other datasets; our new Linked Data will be interlinked with Linked data from Publicspending and with Geonames.org a geospatial registry where already 9 million URIs for place names reside. As a use case of our newly published Linked Data, we will develop a faceted browser that will enable end-users to query all Greek companies residing in our SPARQL endpoint based on various facets.

5.3 Publishing and consumption pipeline

In the next sections we will describe each phase of our pipeline which is based on the Hausenblas Linked Government Data life cycle.

5.3.1 Data awareness

The adoption of OGD for use in the Web of Data depends of its availability and a necessary first step into expanding the Web of Data with government data, is its discovery process. As the first level in our publishing pipeline, we had to identify and select the government data that we considered useful to transform to Linked Data. Our initial purpose was to publish as Linked Data all Greek public entities so our focus was concentrated in finding relevant datasets.

Our search started at Di@vgeia where all public entities should publish their decisions; therefore a full list of all the public administration entities, should be present. The API offered by Di@vgeia uses taxonomy for Organizations and Organization Units in XML but the only property it describes is the organization name while other organization basic elements like Legal Identifier, Organization type and address could not be retrieved.

So, our search expanded in every Greek Website that would have probabilities to offer basic data for all the Greek public entities and simultaneously we were searching for every available dataset that would have common concepts with the entities and could be integrated and produce a meaningful application. The absence of a central portal where all raw datasets of the Greek Government could be made available to the public and the limited datasets that were published in machine-readable formats made the whole procedure time-consuming and discouraging.

Finally the desired dataset was spotted at *Publicspending.gr*. *Publicspending*'s data included all the information we needed not only for Greek public entities but for all Greek

companies, so we decided to describe as Linked Data not only Greek *public* entities but *all* Greek legal entities, public and private. *Publicspending*'s companies data is being obtained directly through the Web services that the General Secretariat of Information Systems of Ministry of Finance has made available to the public, so their up-to-date, unique for each company and they have excellent provenance. *Publicspending* developer's team willingly agreed to give us a file with all the Greek companies they have obtained from the Web services so far, in our desired CSV format. Until now they have retrieved data for 28,000 companies from a total of 300,000. Further arrangements have been made to give us in regular intervals the remaining companies. This dataset would be the basis for the next phase of our pipeline, the phase of modeling.

5.3.2 Modeling

Once the desired datasets have been discovered, the next step in our pipeline should be to model our schema so that it can be integrated in the linked data cloud.

The output of this first phase of our pipeline was a dataset in CSV format obtained from Publicspending's SPARQL endpoint that contained 28,900 Greek companies (not only public entities but also any type of Greek companies). Table 7 explains how each data column of this file was used to describe the legal entities with the Core Vocabularies.

	Publicspending's converted CSV Columns	***Model schema / Core Vocabularies /***		
1	Publicspending's URI based on their domain	Used for interlinking		
2	Name of the company (As appears on TAXIS	*legal name*		
3	Zip code	Core Location	PostCode	Excel transformation to full address
4	Street		Thoroughfare	
5	City		PostName	
6	Street Number		Loc. Designator	
7	Phone	*Not Used*		
8	AFM (The Greek unique identifier of every legal entity)	Used for constructing URIs / *Legal Identifier skos:notation*		
9	Status (With status they actual give the company type)	Used for *company type* skos: Preflabel / Translations and used to construct *company type* URIs		
10	cpaCode (code for describing activities)	Not Used at present. Future work for *company activity*		
11	Doy : the code of the regional tax office that	*Not Used*		
12	Doy Name : the name of the regional tax of-	*Legal Identifier dcterms:creator*		
13	Branches	*Not Used*		
14	Start date (the registration date of the compa-	*Legal Identifier dcterms:created*		
15	End date (if there is any)	Check if exist then *company status* "Inactive" else "Normal Activity"		

Table 7 Description of Publicspending's CSV file

To make the data from the CSV file compatible with properties and classes described with the Core Vocabularies extended Excel transformations were performed so that each column would contain the right format and information. Figure 10 displays part of the CSV file after the performed transformations.

payer	name	zip	street	city	number	number New	phone	afm	status	new status	cpaCode	doy	doyName	startDate Time	startDate	end Date	full address	CompanyStatus
http://public	ΣΥΛΛΟΓΟΣ ΔΙΚΗΓΟΡΙΚΟΣ	54627	ΔΙΚΑΣΤΙΚΟ	ΘΕΣΣΑΛΟΝ	0		500870	090010794	ΣΥΛΛΟΓΟ	assoc	78826	4212	Β ΘΕΣΣΑΛΟΝΙΚ	1901-01-01T0	1/1/1901		ΔΙΚΑΣΤΙΚΟ	NormalActivity
http://public	ΣΥΛΛΟΓΟΣ ΠΟΝΤΙΩΝ ΔΡΑ	66100	19 ΜΑΙΟΥ	ΔΡΑΜΑ	4	4	Null	099448796	ΣΥΛΛΟΓΟ	assoc	94991601	5111	ΔΡΑΜΑΣ	1949-01-31T0	31/1/1949		19 ΜΑΙΟΥ 4	NormalActivity
http://public	ΑΘΛΗΤΙΚΟΣ ΠΟΛΙΤΙΣΤΙΚΟ	15123	ΚΗΦΙΣΙΑΣ	ΜΑΡΟΥΣΙ	37	37	Null	998988090	ΣΥΛΛΟΓΟ	assoc	78894	1135	ΑΜΑΡΟΥΣΙΟΥ	2002-03-05T0	5/3/2002		ΚΗΦΙΣΙΑΣ 3	NormalActivity
http://public	ΘΕΑΤΡΟ ΤΟΥ ΠΑΙΔΙΟΥ	17671	ΕΥΡΥΔΙΚΗ	ΚΑΛΛΙΘΕΑ	9	9	9595084	099075510	ΣΥΛΛΟΓΟ	assoc	90031000	1130	Α ΚΑΛΛΙΘΕΑΣ	1999-11-30T0	30/11/1999		ΕΥΡΥΔΙΚΗΣ	NormalActivity
http://public	ΕΜΠΟΡΙΚΟΣ ΣΥΛΛΟΓΟΣ Σ	50300	ΣΙΑΤΙΣΤΑ	ΣΙΑΤΙΣΤΑ	0		Null	099290927	ΣΥΛΛΟΓΟ	assoc	78877	4543	ΝΕΑΠΟΛΗΣ ΒΟ	1987-04-14T0	14/4/1987		ΣΙΑΤΙΣΤΑ Σ	NormalActivity
http://public	ΠΟΛΙΤΙΣΤΙΚΟΣ ΣΥΛΛΟΓΟΣ	47100	ΠΡΙΟΒΟΛΑ	ΑΡΤΑ	11	11	Null	090369309	ΣΥΛΛΟΓΟ	assoc	94991601	6111	ΑΡΤΑΣ	1980-03-20T0	20/3/1980		ΠΡΙΟΒΟΛΟ	NormalActivity
http://public	ΕΠΑΓΓΕΛΜΑΤΙΚΟ ΣΩΜΑΤΕ	64004	ΘΑΣΟΣ	ΘΑΣΟΣ	0		Null	800125984	ΣΥΛΛΟΓΟ	assoc	78198	5311	ΘΑΣΟΥ	1982-02-16T0	16/2/1982		ΘΑΣΟΣ ΘΑ	NormalActivity
http://public	ΣΥΛΛΟΓΟΣ ΥΠΑΛΛΗΛΩΝ Β	54624	ΤΣΙΜΙΣΚΗ	ΘΕΣΣΑΛΟΝ	29	29	231056	099389860	ΣΥΛΛΟΓΟ	assoc	78826	4212	Β ΘΕΣΣΑΛΟΝΙΚ	1976-03-31T0	31/3/1976		ΤΣΙΜΙΣΚΗ 2	NormalActivity
http://public	ΣΥΛΛΟΓΟΣ ΕΛΕΥΘΕΡΩΝ Ε	62122	ΕΘΝ ΑΝΤΙ	ΣΕΡΡΕΣ	10	10	2,32E+09	999218695	ΣΥΛΛΟΓΟ	assoc	78887	5621	Α ΣΕΡΡΩΝ	1998-04-10T0	10/4/1998		ΕΘΝ ΑΝΤΙΣ	NormalActivity

Figure 10 Final CSV File Format

The updated CSV file contained the data in a format compatible with properties of the Core Vocabularies. Figure 11 gives a graphical representation of the way each data column from the CSV file was used to describe classes and properties of the Core Vocabularies while the whole procedure followed and the modeling of our schema will be described with details in the next paragraphs.

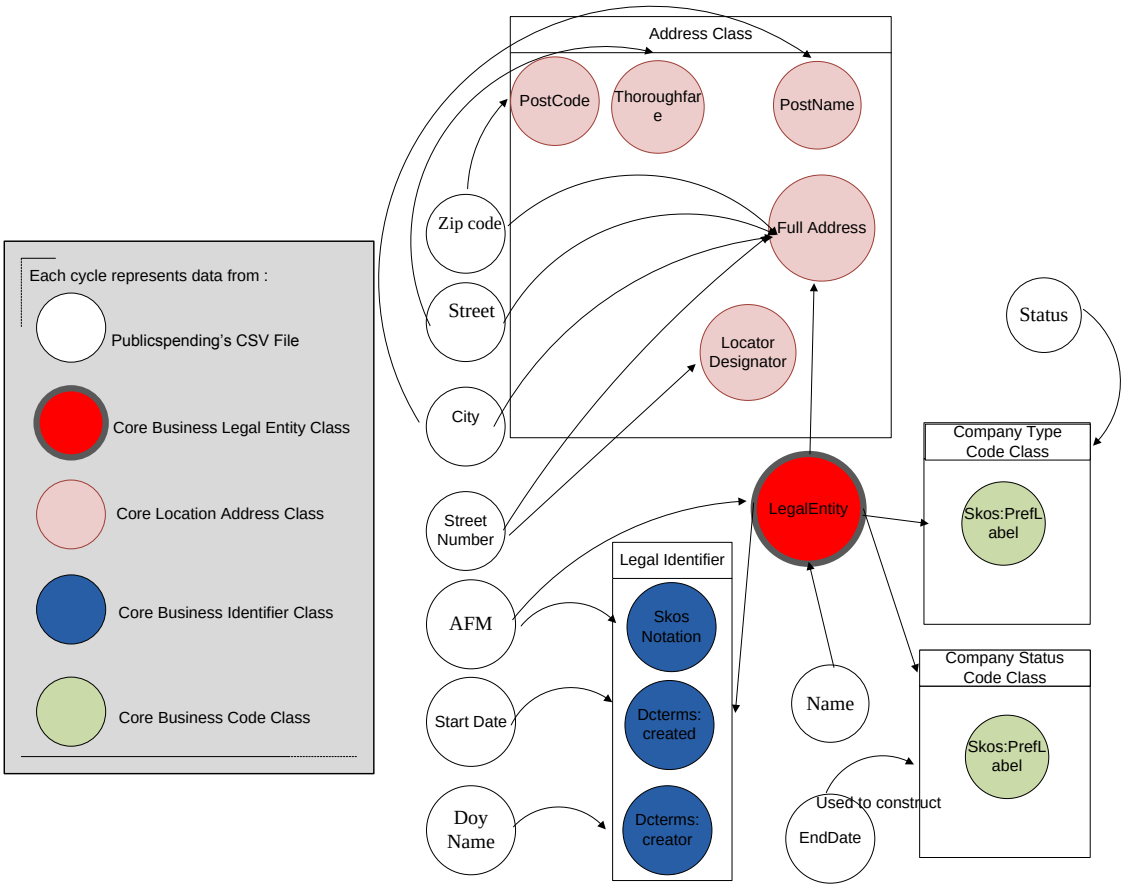


Figure 11 Modeling data from CSV file with the Core Vocabularies

In order to publish data on the Web, these data should be identified by URIs; either they represent real-world entities, abstract concepts or web documents. We use to refer to these data as *resources.* In the current phase, the phase of modeling, the resources in the modelisation problem should first be identified, along with the concepts that best describe them. Since we have decided to model our schema with the Core Business Vocabulary, we could immediately identify our resources and concepts based on the vocabulary's classes and properties:

Legal Entities

In our case study, companies are the key resources and we have decided to use Core Business Vocabulary to describe them in an EU interoperable way. Following the Core Business Vocabulary, companies are defined as *Legal Entities*. Each *Legal Entity* should have one *legal name* which must be the official name of the company as it is registered in the official national tax system. An example is given in Figure 12. Our initial dataset included official company names as it is taken from the General Secretariat of Information Systems of Ministry of Finance through its Web services[65].

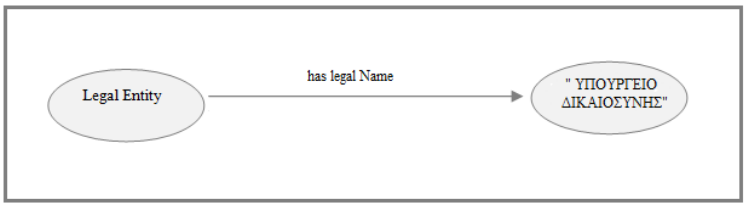


Figure 12 Description of the Legal Name

Legal Identifier

Each Legal Entity must have a unique identifier assigned by the issuing authority within a given jurisdiction. In Greece, the company's tax number (referred to as AFM a 10 digit number) fulfills this need and it is included in our initial dataset.

Legal identifier, as any identifier associated with a legal entity, should not be described with a literal as Figure 13 depicts because it would contain no semantics. Instead with the Core Vocabulary identifiers are represented as a sub-property of the identifier class (Figure 14).

[65] GSIS Web Services http://www.gsis.gr/wsnp/wsnp.html [Accessed October 2012]

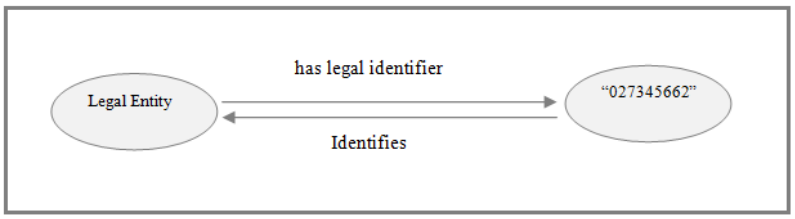


Figure 13 Wrong description of Legal Identifier with no semantics

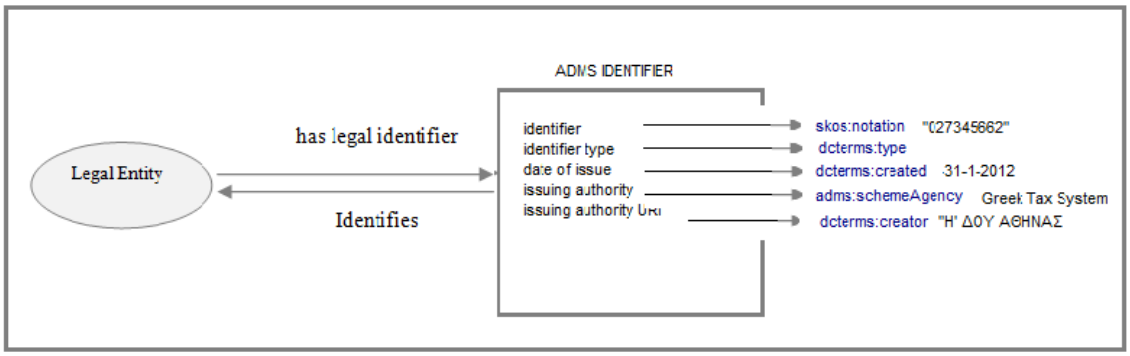


Figure 14 Describing Legal Identifier with the Identifier Class

Company Address

As mentioned in the previous paragraph, every identifier should not be defined as a literal but should be described with the Identifier Class. In the same concept the address of a company should not be a free text but should be described with an appropriate class. The Core Location Vocabulary [45] provides a class describing addresses, so the address of the company can be a sub-property of this class as it is depicted in Figure 15.

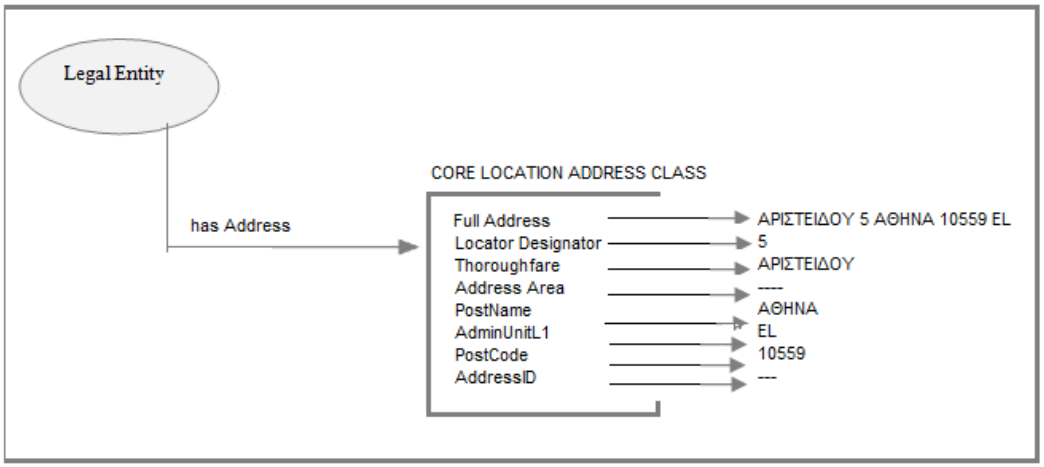


Figure 15 Describing company address with Core Location

Company Type, Company Status, Company Activity

Each company has a type. We have distinguished 66 different types of companies in our dataset. Describing company's type with just a literal (e.g. "ΝΠΔΔ") would not be meaningful to a non-Greek citizen. That is the main reason that concepts *like company*

type, *activity* and *status*, should be described with the *Code* class in the Core Business Vocabulary. Code class is a generic class to describe any kind of code and its terms should correspond to controlled vocabularies. Since in Greece there is no controlled vocabulary to describe Greek company types, we had to define them using the SCOS concept in a compatible with the *Code* class way as it is recommended in the Core Business Vocabulary specification [49]. In the next paragraph we will describe the whole procedure followed:

Each Greek company type was assigned a specific URI constructed based on its initials e.g. **http://linkeddata.ihu.edu.gr/recource/grtypes/npdd** for "**Ν**ομικό **Π**ρόσωπο **Δ**ημοσίου **Δ**ικαίου". Using SKOS we described the Greek term for company type as a PrefLabel, while in a second PrefLabel we tried to correlate them with common business company types or English translations wherever possible, to increase interoperability. With SKOS Notation we described widely used Greek company types initials like "ΝΠΔΔ" in the previous example, if they existed.

The correlation is not complete. However it serves the basic and most common company types and it creates the technical background, so whenever a common EU schema for all company types is agreed, the correlation with existing Greek types can be immediately implemented. Table 8 includes all 66 different Greek company types along with their SKOS descriptions while it reveals the ones that a correlation was not feasible.

URI with prefix http://linkedda-ta.ihu.edu.gr/	Greek Company Type described in first skos PrefLabel	Common English described in the second skos PrefLabel	Skos Notation
npdd	Νομικό Πρόσωπο Δημοσίου Δικαί-ου@gr	Public entity@e	ΝΠΔΔ
ae	Ανώνυμη Εταιρία@gr	Public Limited Company SA@en	ΑΕ
oe	Ομόρρυθμη εταιρεία@gr	General Partnership@en	ΟΕ
ee	Ετερόρρυθμη εταιρεία@gr	Limited Partnership@en	ΕΕ
epe	Εταιρεία περιορισμένης ευθύνης@gr	Limited Liability Company (LLC or Ltd)@en	ΕΠΕ
assoc	Σύλλογος@gr	Association@en	ΣΥΛΛΟΓΟΣ
mepg	Δημοτική Κοινωφελής Επιχείρηση@g	Municipal Enterprice for the public good@en	ΔΗΜΚΟΙΝΕ-ΠΙΧ
consortium	Κοινοπραξία@gr	Consortium@en	*Same as PrefLabel*
ac	ΑΓΡΟΤΙΚΟΣ ΣΥΝΕΤΑΙΡΙΣΜΟΣ@gr	Agro Co-operative@en	*Same as PrefLabel*
bo	ΣΥΝΠΛΟΙΟΚΤΗΣΙΑ@gr	Boat Co-operative@en	*Same as PrefLabel*
npidnp	ΑΛΛΟ ΝΠΙΔ ΜΗ ΚΕΡΔΟΣΚΟΠΙΚΟ@gr	Legal entity governed by private law non profita-ble@en	*Same as PrefLabel*

aspe	ΑΣΤΙΚΟΣ ΣΥΝΕΤΑΙΡΙΣΜΟΣ ΠΕΡΙΟ-ΡΙΣΜ.ΕΥΘΥΝΗΣ@gr	Urban Cooperative Limited Liability@en	*Same as PrefLabel*
ade	ΑΜΙΓΗΣ ΔΗΜΟΤΙΚΗ ΕΠΙΧΕΙΡΗΣΗ@gr	Municipal Corporation@en	*Same as PrefLabel*
kk	ΚΟΙΝΩΝΙΑ ΚΛΗΡΟΝΟΜΩΝ@gr	SOCIETY HEIRS@en	*Same as PrefLabel*
dy	ΔΗΜΟΣΙΑ ΥΠΗΡΕΣΙΑ@gr	Public Service@en	*Same as PrefLabel*
epem	ΕΠΕ ΜΟΝΟΠΡΟΣΩΠΗ@gr	Limited Liability One person Company @en	*Same as PrefLabel*
institution	Ιδρυμα@gr	**Institution**@en	*Same as PrefLabel*
oas	ΟΜΟΣΠΟΝΔΙΑ ΑΣΤΙΚΩΝ ΣΥΝΕΤΑΙΡΙ-ΣΜΩΝ@gr	Urban Cooperative Federa-tion@en	*Same as PrefLabel*
som	ΣΩΜΑΤΕΙΟ@gr	Societies@en	*Same as PrefLabel*
dk	ΔΙΑΧΕΙΡΙΣΗ ΚΤΙΡΙΩΝ@gr	Facility Management@en	*Same as PrefLabel*
npidp	ΑΛΛΟ ΝΠΙΔ ΚΕΡΔΟΣΚΟΠΙΚΟ@gr	Legal entity governed by private law profitable@en	*Same as PrefLabel*
Ya	ΥΠΟΚΑΤΑΣΤΗΜΑ ΑΛΛΟΔΑΠΗΣ ΕΤΑΙΡΙ-Α@gr	Foreign Subsidairy Company@en	*Same as PrefLabel*
amke	ΑΣΤΙΚΗ ΜΗ ΚΕΡΔΟΣΚΟΠΙΚΗ ΕΤΑΙΡΙ-Α@gr	Non-profit organization@en	*Same as PrefLabel*
aeota	ΑΕ ΟΤΑ@gr	Public Municipality Limited Company @en	*Same as PrefLabel*
Aemmf	ΑΛΛΟΔΑΠΕΣ ΕΤΑΙΡΙΕΣ ΜΕΛΗ ΜΗ ΦΠ@gr	Foreign Companies@en	*Same as PrefLabel*
easeas	ΕΝΩΣΗ ΑΓΡΟΤΙΚΩΝ ΣΥΝΕΤΑΙΡΙΣΜΩΝ Ε.Α.Σ.@gr	Agro Union Cooperation@en	*Same as PrefLabel*
kspe	ΚΟΙΝΩΝΙΚΟΣ ΣΥΝΕΤΑΙΡΙΣΜΟΣ ΠΕΡ ΕΥΘΥΝΗΣ@gr	Social Union Limited Liabil-ity@en	*Same as PrefLabel*
aepmk	ΑΛΛΗ ΕΝΩΣΗ ΠΡΟΣΩΠΩΝ ΜΗ ΚΕΡΔΟ-ΣΚΟΠΙΚΗ@gr	Non profitable Union@en	*Same as PrefLabel*
kaso	ΚΟΙΝΟΠΡ.ΑΓΡΟΤΙΚΩΝ ΣΥΝΕΤΑΙΡ. ΟΡ-ΓΑΝΩΣΕΩΝ@gr	Agro Cooperative@en	*Same as PrefLabel*
ask	ΑΣΤΙΚΟΣ ΣΥΝΕΤΑΙΡΙΣΜΟΣ ΚΕΡΔΟΣΚΟ-ΠΙΚΟΣ@gr	Urban Profitable Cooperative@en	*Same as PrefLabel*
ane	ΑΜΙΓΗΣ ΝΟΜΑΡΧΙΑΚΗ ΕΠΙΧΕΙΡΗ-ΣΗ@gr	Regional Business@en	*Same as PrefLabel*
aeed	ΑΣΤΙΚΗ ΕΠΑΓΓΕΛΜΑΤΙΚΗ ΕΤΑΙΡΙΑ ΔΙ-ΚΗΓΟΡΩΝ@gr	Urban Professional Business of Layers@en	*Same as PrefLabel*
Imeepe	ΙΜΕ ΕΠΕ@gr	Not Known	*Same as PrefLabel*
Enasyn	ΕΝΩΣΗ ΑΣΤΙΚΩΝ ΣΥΝΕΤΑΙΡΙΣΜΩΝ@gr	Not Known	*Same as PrefLabel*
Aemon	ΑΕ ΜΟΝΟΠΡΟΣΩΠΗ@gr	Not Known	*Same as PrefLabel*
Kke	ΚΟΙΝΟΤΙΚΗ ΚΟΙΝΩΦΕΛΗΣ ΕΠΙΧΕΙΡΗ-ΣΗ@g	Not Known	*Same as PrefLabel*
aenp	ΑΕ ΜΗ ΚΕΡΔΟΣΚΟΠΙΚΗ@g	Not Known	*Same as PrefLabel*
aekps	ΑΕ ΕΠΙΧ ΠΡΟΓΡΑΜ ΚΠΣ (Ν.2860/2000)@g	Not Known	*Same as PrefLabel*
ne959	ΝΑΥΤΙΚΗ ΕΤΑΙΡΙΑ Ν.959/79@gr	Not Known	*Same as PrefLabel*
kasdk	ΚΟΙΝΩΝΙΑ ΑΣΤΙΚΟΥ ΔΙΚΑΙΟΥ ΚΕΡΔΟ-ΣΚΟΠΙΚΗ@gr	Not Known	*Same as PrefLabel*

dse	ΔΗΜΟΤΙΚΗ ΣΥΝΕΤΑΙΡΙΣΤΙΚΗ ΕΠΙΧΕΙΡΗ-ΣΗ@gr	Not Known	*Same as PrefLabel*
aepk	ΑΛΛΗ ΕΝΩΣΗ ΠΡΟΣΩΠΩΝ ΚΕΡΔΟΣΚΟ-ΠΙΚΗ@g	Not Known	*Same as PrefLabel*
ade	ΑΜΙΓΗΣ ΔΙΑΔΗΜΟΤΙΚΗ ΕΠΙΧΕΙΡΗ-ΣΗ@gr	Not Known	*Same as PrefLabel*
Aemsese	ΑΛΛΟΔ.ΕΤΑΙΡΙΑ ΜΕ ΣΥΜΒ.ΕΡΓΟΥ ΣΤΗΝ ΕΛΛΑΔΑ@gr	Not Known	*Same as PrefLabel*
nepa	ΝΕΠΑ ΝΑΥΤΙΛ ΕΤΑΙΡ ΠΛΟΙΩΝ ΑΝΑΨΥΧ Ν3182/03@gr	Not Known	*Same as PrefLabel*
eemhfpae	ΕΝΤΟΛΕΑΣ ΕΕ ΜΗ ΦΠ ΑΕ@gr	Not Known	*Same as PrefLabel*
ake	ΑΣΤΙΚΗ ΚΕΡΔΟΣΚΟΠΙΚΗ ΕΤΑΙΡΙΑ@gr	Not Known	*Same as PrefLabel*
aeia	ΑΛΛΟΔΑΠΗ ΕΤΑΙΡΙΑ ΙΔΙΟΚΤΗΣΙΑΣ ΑΚΙ-ΝΗΤΟΥ@gr	Not Known	*Same as PrefLabel*
8hoee	8Η ΟΔΗΓΙΑ ΕΕ@gr	Not Known	*Same as PrefLabel*
Annaet	ΑΝ 378/68 ΚΑΙ Ν 27/75 ΝΑΥΤΙΛΙΑΚΕΣ ΕΤΑΙΡ@gr	Not Known	*Same as PrefLabel*
pkes	ΠΟΛΙΤΙΚΑ ΚΟΜΜΑΤΑ ΕΚΛΟΓΙΚΟΙ ΣΥΝ-ΔΥΑΣΜΟΙ@gr	Not Known	*Same as PrefLabel*
Aeaoed	ΑΕ ΟΑΕΔ (Ν.2956/2001)@gr	Not Known	*Same as PrefLabel*
aemk	ΑΣΤΙΚΟΣ ΣΥΝΕΤΑΙΡΙΣΜΟΣ ΜΗ ΚΕΡΔΟ-ΣΚΟΠΙΚΟΣ@gr	Not Known	*Same as PrefLabel*
eper	ΕΠΙΤΡΟΠΗ ΕΡΑΝΩΝ@gr	Not Known	*Same as PrefLabel*
asape	ΑΣΤΙΚΟΣ ΣΥΝΕΤΑΙΡΙΣΜΟΣ ΑΠΕΡΙΟ-ΡΙΣΤ.ΕΥΘΥΝΗΣ@gr	Not Known	*Same as PrefLabel*
etan89	ΕΤΑΙΡΙΑ Α.Ν. 89/67@gr	Not Known	*Same as PrefLabel*
eeemdpepe	ΕΝΤΟΛΕΑΣ ΕΕ ΜΗ ΦΠ ΕΠΕ@gr	Not Known	*Same as PrefLabel*
Do	ΔΙΕΘΝΗΣ ΟΡΓΑΝΙΣΜΟΣ@gr	Not Known	*Same as PrefLabel*
an89hmae	ΑΝ 89/67 ΗΜΕΔΑΠΗ ΑΕ@gr	Not Known	*Same as PrefLabel*
akoine	ΑΜΙΓΗΣ ΚΟΙΝΟΤΙΚΗ ΕΠΙΧΕΙΡΗΣΗ@gr	Not Known	*Same as PrefLabel*
imeepem	ΙΜΕ ΕΠΕ ΜΟΝΟΠΡΟΣΩΠΗ@gr	Not Known	*Same as PrefLabel*
adiaep	ΑΜΙΓΗΣ ΔΙΑΚΟΙΝΟΤΙΚΗ ΕΠΙΧΕΙΡΗ-ΣΗ@gr	Not Known	*Same as PrefLabel*
kadmk	ΚΟΙΝΩΝΙΑ ΑΣΤΙΚΟΥ ΔΙΚΑΙΟΥ ΜΗ ΚΕΡ-ΔΟΣΚΟΠΙΚΗ.@gr	Not Known	*Same as PrefLabel*
aees	ΑΣΤΙΚΗ ΕΠΑΓΓΕΛΜ ΕΤΑΙΡΙΑ ΣΥΜΒΟ-ΛΑΙΟΓΡΑΦΩΝ@gr	Not Known	*Same as PrefLabel*
an89alet	ΑΝ 89/67 ΑΛΛΟΔΑΠΕΣ ΕΤΑΙΡΙΕΣ@gr	Not Known	*Same as PrefLabel*
ksyne	ΚΟΙΝΟΤΙΚΗ ΣΥΝΕΤΑΙΡΙΣΤΙΚΗ ΕΠΙΧΕΙΡΗ-ΣΗ@gr	Not Known	*Same as PrefLabel*

Table 8 Greek Company Types described with SKOS

As for the concepts of company status and company activity they should be described by following the same procedure: declaring the classes Company Status and Company Activity and describing the concept schema using terms from SKOS vocabularies.

Company status was not included in the original dataset but we assumed normal activity to every company except the ones that had *end date* in the original dataset. In that case we considered these companies to be *Inactive*. This assumption provided us the opportunity to create the mechanisms so when we obtain company status for all legal entities, implementing it would be a matter of simple substitutions.

Figure 16 describes the modeling while Figure 17 depicts a sample of how *company type and company status* are implemented with the new classes and schemas. Their description with RDF followed a hand-authoring procedure because the file was small, translations had to be made and there weren't many repeating patterns. Core Vocabularies specification [44] recommends to include these RDF descriptions in the same RDF file with the main dataset. We decided, at least temporarily, to keep it separate, due to the huge bulk of triples describing our main dataset and the fact that regularly new versions of this dataset should be transformed to RDF. Appending the definition of the classes and schema in every new RDF file wouldn't be a very good practice. When the number of companies and the company types that describe them is standardized, the two RDF files may easily become one.

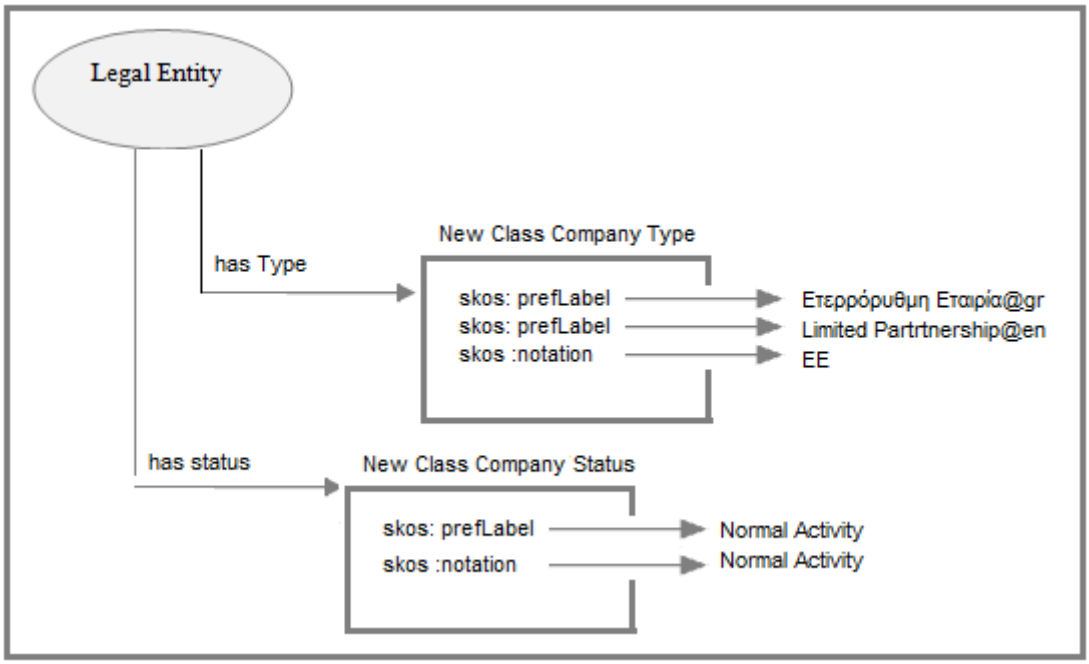


Figure 16 Modeling company types and status

The file's name is *CompanyTypes.rdf* and it was validated using *W3C validator*. A part of the file, declaring the type class and a specific company type, is described in Figure 17:

```
!--Declare the class and the concept schema for Company Type -->
<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/def/code/grtypes">
<rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-
            schema#Class"/>
<skos:prefLabel> Greek company types</skos:prefLabel>
<rdfs:comment>This is the class of Greek company types</
rdfs:comment>
<rdfs:subClassOf >http://www.w3.org/2004/02/skos/core#Concept</
rdfs:subClassOf >
</rdf:Description>

<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/resource/grtypes/
npdd">
<rdf:type rdf:resource="http://www.w3.org/2004/02/skos/
core#ConceptScheme"/>
<rdf:type rdf:resource="http://linkeddata.ihu.edu.gr/def/code/grtypes"/>
<skos:prefLabel>Νομικό Πρόσωπο Δημοσίου Δικαίου@gr</
skos:prefLabel>
<skos:prefLabel>Public entity@en</skos:prefLabel>
<skos:notation>ΝΠΔΔ</skos:notation>
</rdf:Description>
```

Class Declaration

Declaration of a single company type - ΝΠΔΔ

Figure 17 New controlled vocabulary for company type and status

Figure 18 depicts our complete modelisation schema that includes all the resource and the terms that we will use to describe them.

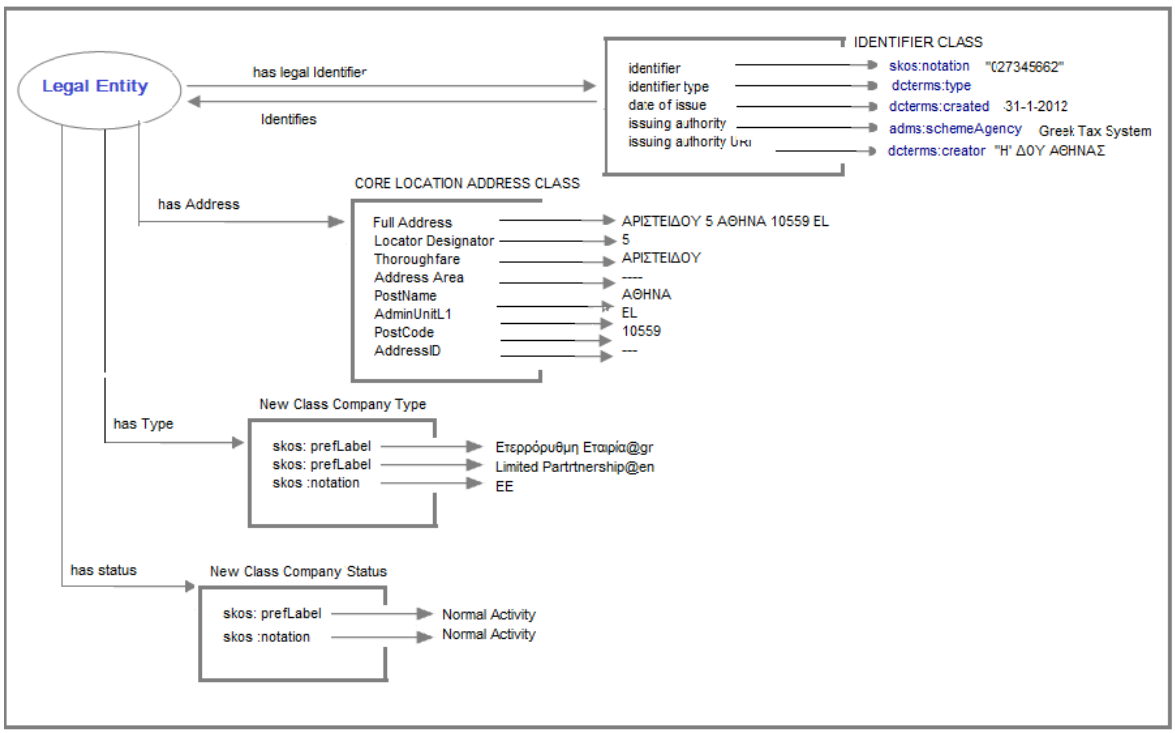


Figure 18 Final Conceptualization model

Designing the URI schema

We have identified *Legal entities*, *address*, *legal identifiers* as the main resources in our schema. In the Web of Data each resource should be identified by a URI. Designing the URI schema for each kind of resource will complete the phase of modeling.

As the first Linked Data principle suggests, each resource should be represented by a URI. Furthermore, URI should be http-based and *dereferenceable* so machines and people can look up the resources. [37, 50] If people choose to see a URI through a web browser then an html description of the resource must be returned. If it is a Linked data client that requests the URI, then an RDF description of the resource should be returned. This can be achieved with an HTTP mechanism called "content negotiation". Through this mechanism when a client sends a message requesting a resource it includes in the *accept header* the kind of document it prefers.

In the Web of Data there is also another issue that should be taken into consideration: a resource can represent real-world concepts. In these cases the URIs of the resource must be unambiguous. For example, one URI should represent the company as a real-world concept and another URI the web document that *describes* this company. There are two different kinds of strategies to deal with this ambiguity and make URIs representing real-world objects dereferenceable: *303 URI* and *Hash URI*

303 URI

In this case, when a server accepts a request for a URI other than a web document, it sends back to the client a *303 See other* message along with the URI of the web document that describes that resource. In a second step, the client retrieves that web document.

Let's see an example with a URI representation of a company named "Orange". We need three URIs. For example:

1. http://example.com/companies/Orange to represent the real-world company
2. http://example.com/companies/Orange.html for the html document describing "Orange"
3. http://example.com/companies/Orange.rdf for the RDF document describing "Orange"

When a client requests the URI of the real-world company, if it is a web browser the server redirects him to http://example.com/companies/Orange.html, and if it is a Linked Data client it redirects him to http://example.com/companies/Orange.rdf that contains the RDF description of the real-world company.

Normally, the html and the RDF documents contain the description of many similar resources. In some cases, if the files are too big, this strategy is recommended as it enables through its two-step mechanism, the retrieval of the specific resource only and not the retrieval of the whole file.

Hash URI

Through this mechanism a hash (#) is appended to represent a non-document resource.

In our previous example, the real-world concept of the "Orange" company would be represented by the following URI: http://example.com/companies#Orange

When the client wants to retrieve the real world object it sends this URI to the server. The server strips off the part that is after the hash and it sends back to the client the whole html or the RDF file without the need of any redirection.

This mechanism avoids the two-step retrieval from the clients but has the disadvantage of returning the RDF descriptions of all the resources that are contained in the same RDF file. It should be usually preferred in small and stable files like the ones describing vocabularies.

In our case study, we describe 28,000 companies. The resulted RDF file is too big to choose the Hash URI method. We shall design the URI schema for our resources based on the 303-URI strategy taking into consideration the recommendations accompanying Core Business Vocabulary [48], as well as recommended good practices that promote simplicity, stability and manageability [49].

Our domain name that would be the basis for the design of the Uri system is *http://linkeddata.ihu.edu.gr* since our triples will be hosted on a server at the International Hellenic University.

1. Legal Entities

   In case of Greece AFM is a unique representation of each company in her national tax system, so it presents as the adequate candidate for the Legal Entity's URI

   ```
   http://linkeddata.ihu.edu.gr /resource/company/{afm}
   ```

   e.g. http://linkeddata.ihu.edu.gr /resource/company/023456745

2. Legal Identifier

   To visually connect legal identifier with legal entity we will use a similar schema

   ```
   http://linkeddata.ihu.edu.gr /resource/company/li{afm}
   ```

e.g. http://linkeddata.ihu.edu.gr /resource/company/li023456745

3. Address

   Similar to legal identifier, but instead of {li} we will use {ra}

http://linkeddata.ihu.edu.gr /resource/company/ra{afm}

   e.g. http://linkeddata.ihu.edu.gr /resource/company/ra023456745

4. Company Type and Company Status.

   In these cases, we should define not only the concept schema but also the classes. We decided, based on Core Vocabularies recommendations to use the following URIs to describe the classes themselves:

http://linkeddata.ihu.edu.gr /def/code/status	Definition of company status class
http://linkeddata.ihu.edu.gr /def/code/grtypes	Definition of company types class

   while for the concept schema:

http://linkeddata.ihu.edu.gr /resource/status/{status}
http://linkeddata.ihu.edu.gr /resource/grtypes/{grtype}

   e.g. http://linkeddata.ihu.edu.gr /resource/grtypes/ae

5.3.3 Publishing

Through the modeling phase we have identified our resources, the concepts that describe them with terms from CompaniesTypes.rdf and Core Vocabularies and we have decided on the URI schema that we will follow. We are now ready to proceed with the publishing phase.

First step in the publishing process is to use the Resource Description Framework (RDF) to convert our resources and the terms that describe them, in the form of triples as described in section 2.2.1. A variety of software tools have been developed for the "RDFizing" of the data. In section 3.3.1 we explained that the selection of the adequate software tool depends upon the nature of the original data. Our dataset is in CSV format, so we have decided to use Google Refine RDF extension to create our basic RDF file, mainly for two main reasons:

1. It uses a very convenient GUI interface to construct the RDF skeleton
2. It provides the capability to export the RDF skeleton and apply it again for similar files. Since we have decided with Publicspending to send us regularly new CSV files, rewriting the RDF skeleton each time would be time-consuming. With Google Refine RDF extension the whole procedure (inserting a new CSV

file with more companies, applying the RDF skeleton template and exporting the RDF file) can be completed within a few minutes.

An overview of the processes followed to construct the basic businesses RDF file based on the facilities that Google Refine RDF Extension offers will follow:

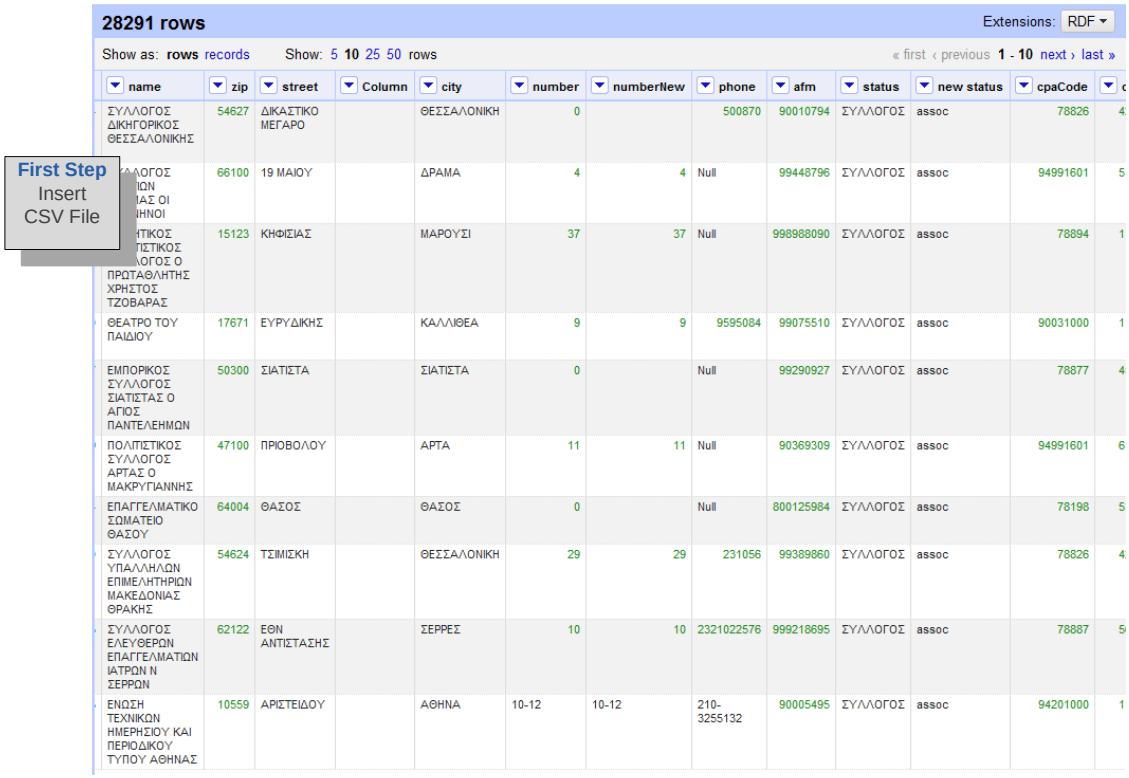


Figure 19 Create the project with Google Refine RDF Extension

As a first step the CSV file needed to be inserted into Google Refine and creation of a new project was in order as seen in Figure 19.

As a second step, the GUI interface was used to create the basic RDF skeleton. Google Refine RDF extension supports the import of vocabularies and use of corresponding prefixes, so we imported the two Core vocabularies and define the prefixes *legal* and *locn*. Through the GUI interface we edit the RDF skeleton by selecting properties from the Core vocabularies and from the controlled vocabulary *CompanyTypes.rdf* (created in the phase of modeling) as depicted in Figure 20.

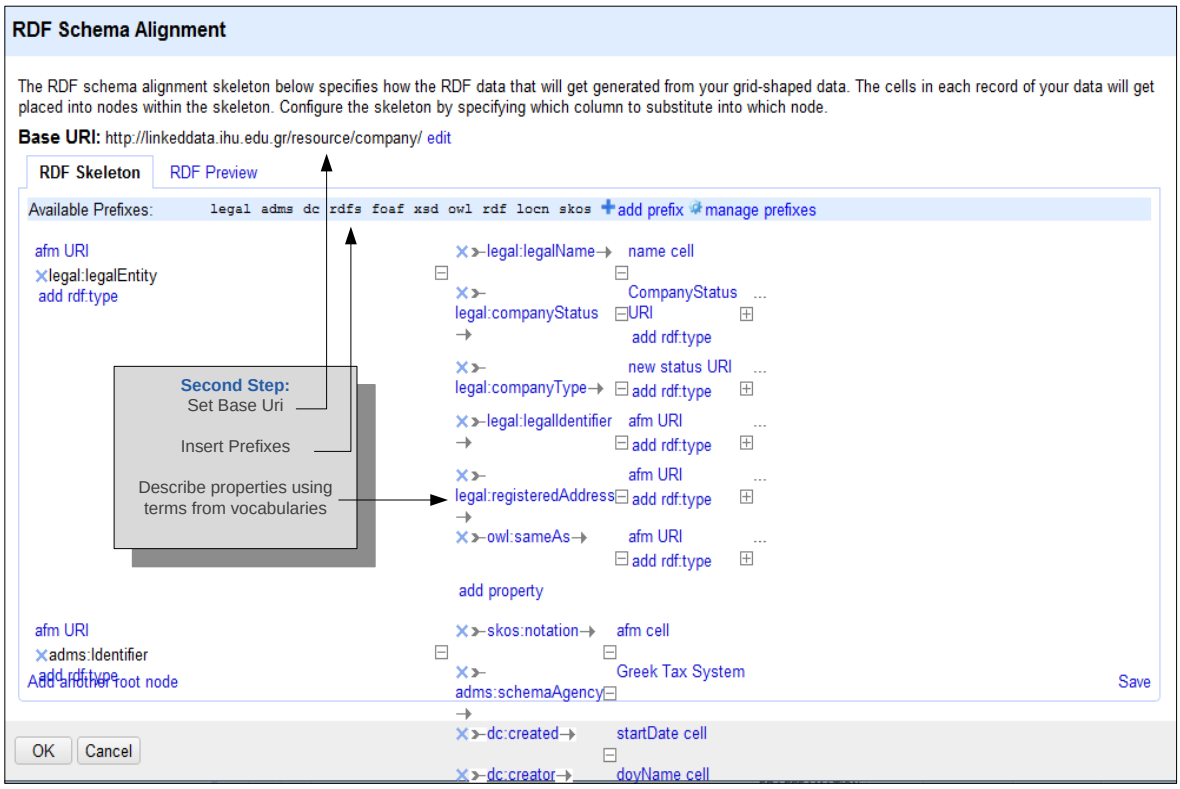


Figure 20 Edit RDF Skeleton with Google Refine

Whenever needed, GREL a simple scripting language that Google Refine offers was used, especially for constructing pattern URIs for the resources, resolved against the base URI. In Figure 21 the construction of the URI pattern for the resource Legal Identifier is illustrated.

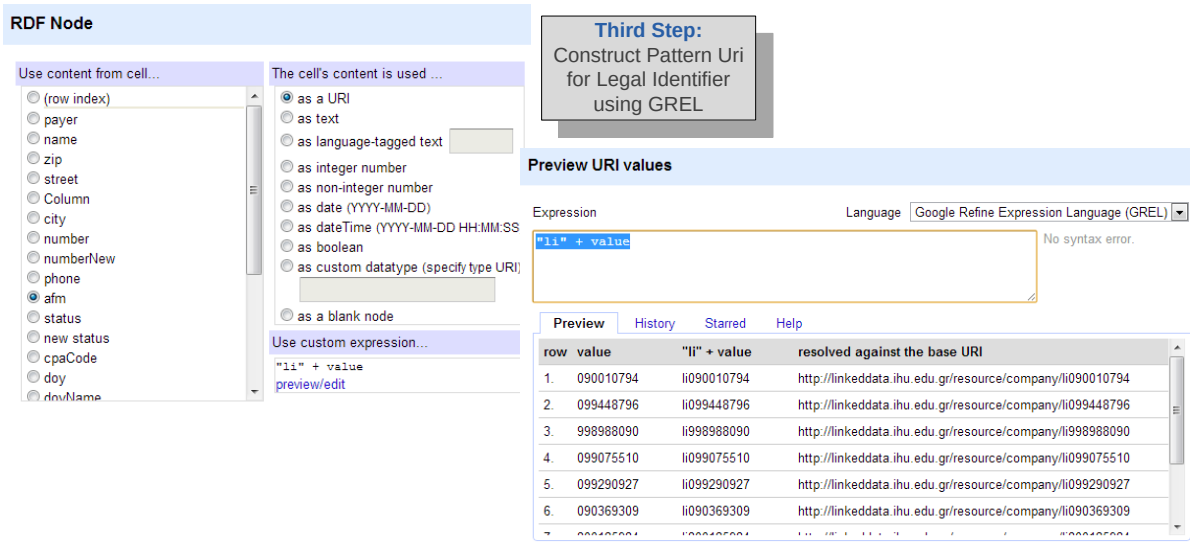


Figure 21 Constructing URI patterns with GREL

Finally the project was exported in RDF/XML serialization as *Companies.rdf*. Through Figure 22 we can see the description of a single company in this file.

Figure 22 RDF description of a legal entity

Hosting and Serving

The primary goal when publishing Linked Data is to make their URIs *dereferenceable*, so they can be discovered. Additionally, some Linked Data publishers store their whole RDF file as *RDF dump* so that it can be downloaded, while the most common procedure is to upload their data on Triple Stores with SPARQL endpoints so they can be retrieved through queries.

Table 3 presents a number of large Triple Stores and in section 3.3.2 we referred to a number of criteria to choose the right store per case. We have stated that the choice depends on:

- Scalability: How many triples can it store and how many triples do you expect to have in the future.

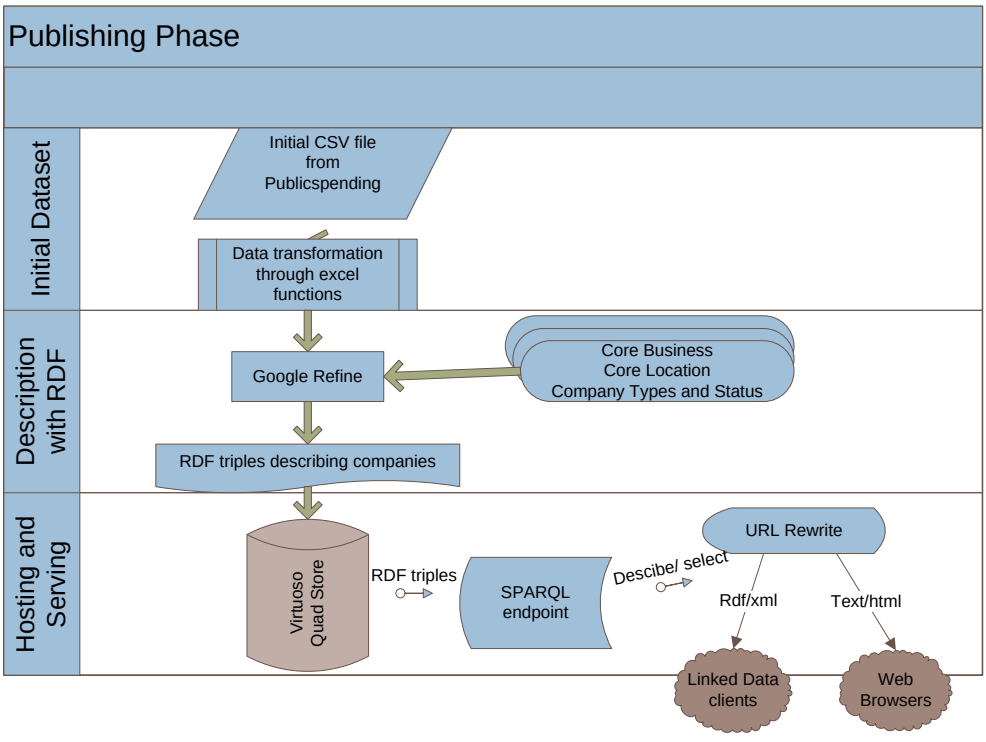
  In our case study, we have 17 triples describing a single company. At present we have 28,800 companies but we expect 300,000 companies in the future, so we must choose a triple store that can handle: 17 * 300,000 = 5,100,000 triples.

- Performance: Compare the performance of Triple stores based on a specific benchmark. We used the *Berlin SPARQL Benchmark* (BSBM) version 3 that has announced its result in the LOD2 stack[66].
- Open source or Commercial: We limit our choices to free Triple Stores.
- Support and community: Wide and active community support would be the natural choice.
- Implementation and facilities that each triple store offers.

Our choice was further guided by the fact that we would like to deploy a faceted browser that works on SPARQL endpoints and was successfully tested on Fuseki, Talis platform and Virtuoso.

Taking all the above into consideration, our final choice was Virtuoso that scales up to billions of triples, supports a SPARQL endpoint, has gone well in the Explore use case during the Berlin Benchmark test cases, is available via an open source license, is relatively easy to set up and has mechanisms to make our URIs dereferenceable.

Shape 23 illustrates the whole procedure from uploading the RDF files that contains our triples to Virtuoso's Quad Store, make them available through Virtuoso's SPARQL endpoint and finally make their URI's dereferenceable by using the URL Rewrite mechanism that Virtuoso provides.



[66] http://lod2.eu/BlogPost/234-berlin-sparql-benchmark-version-3-and-benchmarking-results.html

Figure 23 Hosting and Serving procedures with Virtuoso

Virtuoso offers a user interface called Conductor, through which the RDF files can be uploaded to the triple store (see Figure 24). Actually it is called *quad store* because besides the triplet (subject – predicate – object), Virtuoso adds a fourth parameter to represent each triple's graph. Our basic RDF file that contained the triples describing legal entities and the RDF file containing definition of classes and schema for company types were uploaded to the http://linkeddata.ihu.edu.gr/DAV graph which is the default.

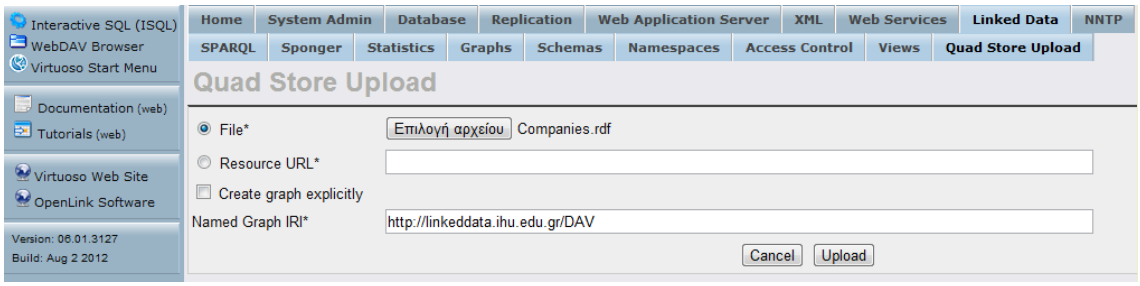


Figure 24 Uploading RDF file into Virtuoso Quad Store

After uploaded in Virtuoso's quad store, data are available through Virtuoso's SPARQL endpoint. A simple query for the retrieval of a single legal entity is described in Figure 25.

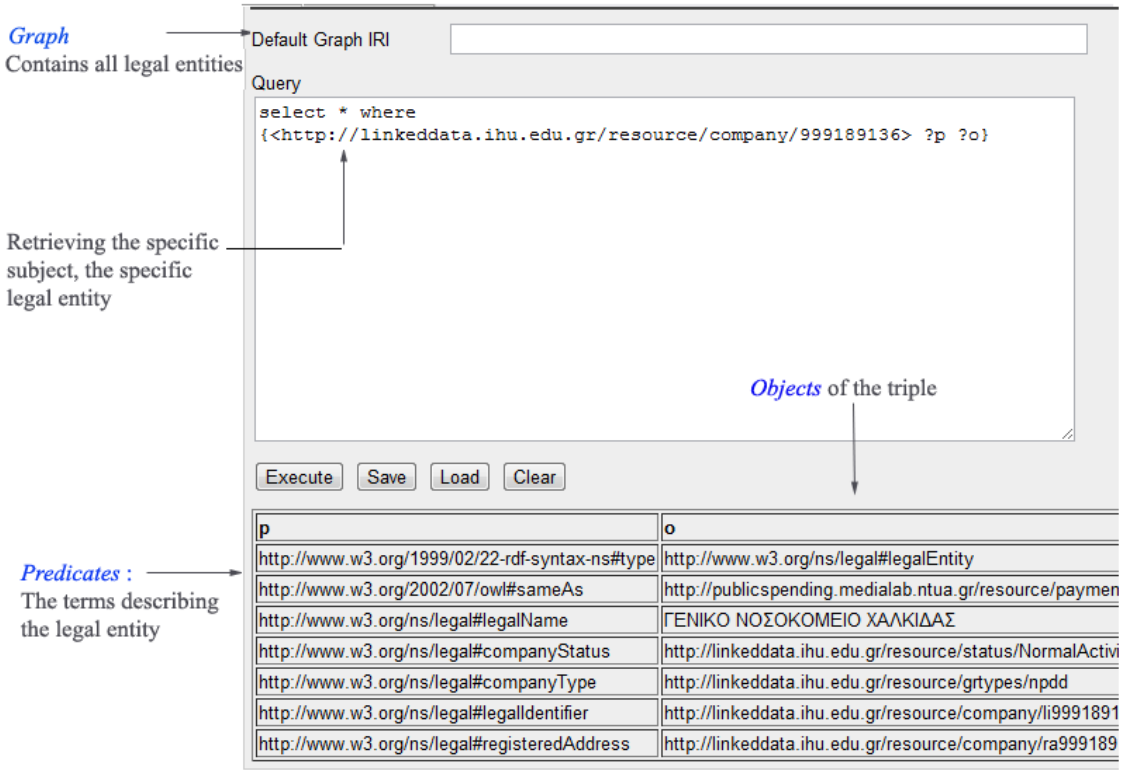


Figure 25 Retrieval of a specific legal entity

As depicted, the objects of the triple are also URIs that can also be accessed and fully described by following the links. Our SPARQL endpoint is available at http://linkeddata.ihu.edu.gr/sparql and there reside all our legal entities as declared by SPARQL feature Count:

```
SELECT (count(?legalentities) as ?count) where
{?legalentities a <http://www.w3.org/ns/legal#legalEntity>}
ORDER BY DESC(?count)
```

count
28291

Dereferencing URIs with Virtuoso

In section 5.4.2 we explained that each resource should have a URI and we described the URI schema we will follow according to 303 URI strategy. We also stated that URIs should be *dereferenceable* according to Linked Data principles. That means that each time a URI is requested from a client, a description of this URI should be available. If it is a web browser that requests the URI, then an HTML file describing the resource should be returned or if it is a Linked data client then an RDF file should be available. Instead of creating physical the HTML and RDF files to describe our resources, Virtuoso offers a mechanism called URL-Rewriter. Through this mechanism, for every request Virtuoso creates "on-the-fly" an HTML or RDF response with data taken from then endpoint and returns it to the client along with *200 OK messages.*

URL-mechanism works based on rules; each rule matches a *source pattern*, the part of the URI that doesn't include the domain name. For example for dereferencing legal entities that have URIs with the following format:

http://linkeddata.ihu.edu.gr/resource/company/{afm}

rules should be created for the pattern /resource/company/

We have distinguished in our URI schema 4 patterns described in Table 9.

URI schema	*Pattern*
http://linkeddata.ihu.edu.gr/**resource/company**/{afm} http://linkeddata.ihu.edu.gr/**resource/company**/li{afm} http://linkeddata.ihu.edu.gr/**resource/company**/ra{afm}	/resource/company
http://linkeddata.ihu.edu.gr/**resource/grtypes**/{type}	/resource/grtypes

http://linkeddata.ihu.edu.gr/**resource/status**/{status}	/resource/status
http://linkeddata.ihu.edu.gr/**def/code**/{status/type}	/def/code

Table 9 URI patterns in our schema

Normally, there would be two rules for each pattern; one to return HTML files and one to return RDF files. The URL-Rewrite rule when a Linked Data client requests a specific legal entity is described in Figure 26 while a similar rule would be created to return HTML files (Figure 27).

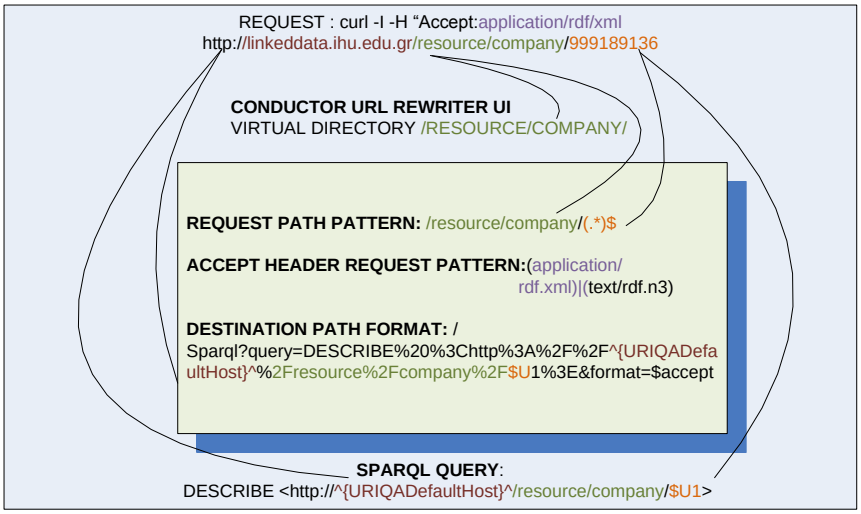


Figure 26 URL Rewriter for an RDF/XML request

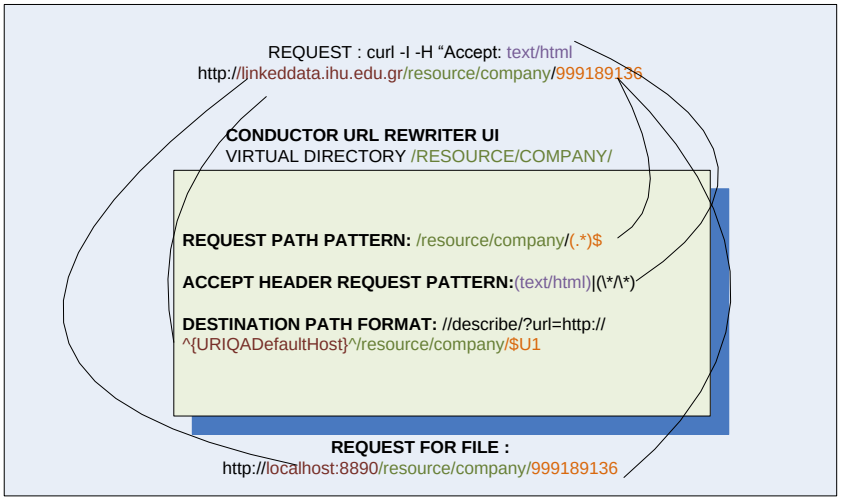


Figure 27 URL Rewriter for an HTML request

To validate our rules, we dereferenced our URIs with a web client, for example a browser, to make sure the proper HTML description is returned from Virtuoso. As Figure 28 depicts for a request Legal Entity, a full description is returned. All the objects of the triples are also URIs so they can also be dereferenced and carry semantics. In Figure 29 we present the Linked Data that are retrieved by following the link of the registered

address of the legal entity while in Figure 30 the description of the legal identifier is depicted

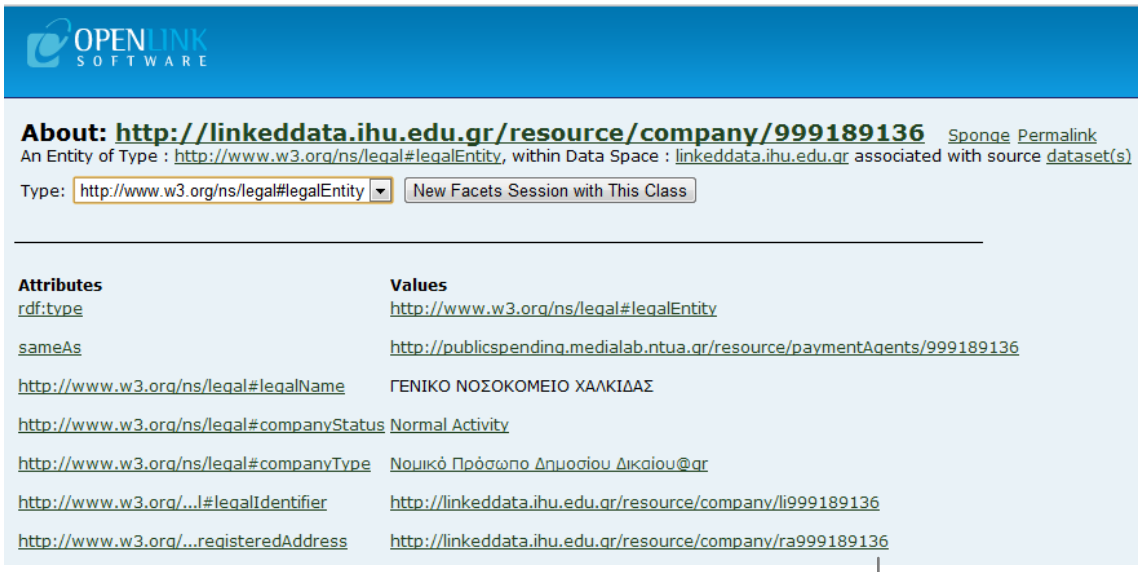


Figure 28 Request for a specific Legal Entity

Figure 29 Request for the Registered Address of the Legal Entity

Figure 30 Request to see the Legal Identifier of the Legal Entity

To validate that the proper RDF description is returned when a Linked Data client requests it, we used *cURL*, a command line tool for transferring files to or from a URL (Figure 31) and set the Accept Header to be N-triples or RDF/ XML.

```
C:\Users\Νατάσα\Downloads\curl_726_0>curl -I -H "Accept: text/rdf+n3" http://loc
alhost:8890/resource/company/90010794
HTTP/1.1 200 OK
Server: Virtuoso/06.01.3127 (Win32) i686-generic-win-32
Connection: Keep-Alive
Date: Thu, 30 Aug 2012 17:37:20 GMT
Accept-Ranges: bytes
Content-Type: text/rdf+n3; charset=UTF-8
Content-Length: 842


C:\Users\Νατάσα\Downloads\curl_726_0>curl -I -H "Accept: application/rdf+xml" ht
tp://localhost:8890/resource/company/90010794
HTTP/1.1 200 OK
Server: Virtuoso/06.01.3127 (Win32) i686-generic-win-32
Connection: Keep-Alive
Date: Thu, 30 Aug 2012 17:37:30 GMT
Accept-Ranges: bytes
Content-Type: application/rdf+xml; charset=UTF-8
Content-Length: 837


C:\Users\Νατάσα\Downloads\curl_726_0>
```

Figure 31 Validation with cURL

As a conclusion, in the current phase we created 2 RDF files. One named *Company-Types.rdf* through a handwriting procedure to declare the classes and the schema of company types and status. The second file, named *Companies.rdf*, contained the description of the legal entities themselves and was created from Publicspending's CSV file with Google Refine RDF Extension. To serve the triples the two files contained though the Web, we uploaded them to Virtuoso's Quad Store, made them available through Virtuoso's SPARQL endpoint and dereference the URI's through Virtuoso's URL-Rewriter mechanism. We are ready now to proceed to the next phase, the phase of discovery related datasets.

5.3.4 Discovery

Using Semantic Web standards like RDF and SPARQL to describe and publish datasets are an important step towards the Web of Data. However, its full potential will be revealed with the discovery of related datasets and their interlinking or else we will be left with small islands of well-structured data.

In this direction, we should first make our data discoverable, so that ingoing links would be created from other data publishers.

Discovery of our Linked Data

The discovery of our data from other RDF users or applications can mainly be achieved through our SPARQL endpoint http://linkeddata.ihu.edu.gr/sparql. For example the following command will retrieve all businesses along with the concepts and data that describe them:

```
Select * where {?s  a <http://www.w3.org/ns/legal#legalEntity>; ?p ?o}
```

However, sometimes it is preferable to share the entire RDF file. In this way data publishers but also end-user application developers would have immediate access to the desired linked data and save time and resources. CKAN [18] is a well-known public data hub specially designed to facilitate data discovery from other publishers based on keywords and its functionality is also provided through a restful API. An extension of CKAN also enables hosting of datasets, updates and keeps track of changes, versions and author informations. To take advantage of these benefits, we uploaded our RDF files in CKAN in a new package called *Greek-legal-entities* with URL http://datahub.io/dataset/greek-legal-entities. All our datasets are available and fully described in this package and can be downloaded under open license.

A VoID file was also created and uploaded to our package in CKAN to act as a bridge between us and the users of our RDF data. The VoID [51] vocabulary is especially designed to express metadata about datasets, access and structural metadata, and links between datasets. A description of our datasets with the VoID vocabulary will provide the users of our data with useful information. Our VoID file *void.ttl* was created by following a handwriting procedure in Turtle syntax and was validated and transformed to the desired RDF/XML syntax to be compatible with our other files with the *rdf:about*[67]

[67] RDFabout validator http://www.rdfabout.com/demo/validator/ [Accessed October 2012]

validator. A code sample of the VoID file in the more readable Turtle syntax that describes the companies' dataset is listed below:

```
<http://linkeddata.ihu.edu.gr/void.ttl>
    a void:DatasetDescription ;
    dcterms:title "A VoiD Description of the Greek Legal Entities Dataset" ;
    dcterms:creator :Varitimou_Natasa ;
    .
<http://linkeddata.ihu.edu.gr/resource/company>
   a void:Dataset;
   foaf:homepage <http://linkeddata.ihu.edu.gr:8080/rdf/browser>;
   rdfs:label "Greek Legal Entities described with the Core Vocabularies";
   dcterms:title "Greek Legal Entities";
  dcterms:description " Greek legal entities described with Core Business and
Core Location Vocabularies";
   dcterms:publisher :Varitimou_Natasa;
   dcterms:created "2012-10-02"^^xsd:date;
   void:feature <http://www.w3.org/ns/formats/RDF_XML>;
   void:exampleResource http://linkeddata.ihu.edu.gr/resource/company/090169674>;
   void:vocabulary <http://www.w3.org/ns/legal>,<http://www.w3.org/ns/locn>;
    void:sparqlEndpoint <http://linkeddata.ihu.edu.gr/sparql/>;
```

A description of our package and our resources in CKAN is depicted in Figure 32.

Figure 32 Greek Legal Entities package in CKAN

Discovery of related Linked Datasets for interlinking

As far as the creation of outgoing links is concerned, four categories of software tools exist that could facilitate our discovery of related data as described in section 3.4. In our case, the use of such tools did not prove very useful and we followed a more based-on-experience approach due to fact that our Linked Data concern Greek datasets and Linked Data movement is still in its infancy in Greece. The quantity of published Linked Data is an important factor, not only for their interlinking but also for the development of end-user applications. As more and more Linked Data are being published, more joint parts can be discovered between datasets that can facilitate interlinking and developing innovative client applications with real-world users.

Through the previous phases of the pipeline we have described Greek public entities and companies as Linked data using data taken from the Publicspending's endpoint. Actually Publicspending's ontology includes the concepts of companies and organizations and a vast variety of properties describing them. Interlinking their resources and ours, which are the same but are described with the Core Vocabularies, would be proven extremely useful. The integration process will be described in section 5.3.5.

Linking our data with big data hubs in the Linked Data cloud will further contribute to the discovery of our data from Linked Data crawlers and browsers.

DBpedia[68] is considered to be a big hub, however its Greek version[69] is not complete and the lack of contents similar to our concepts was apparent. Our efforts to reconcile the names of our legal entities with resources describing companies and organizations from Greek DBpedia (through Google Refine RDF extension reconciliation services) as depicted in Figure 33 had almost zero results. As we will state in our conclusions (chapter 6), the enrichment of Greek DBpedia's content will mainly contribute to the development of Greek Linked Data movement.

[68] http://dbpedia.org and http://dbpedia.org/sparql for the sparql endpoint [Accessed September 2012]

[69] http://el.dbpedia.org and http://el.dbpedia.org/sparql for the sparql endpoint [Accessed September 2012]

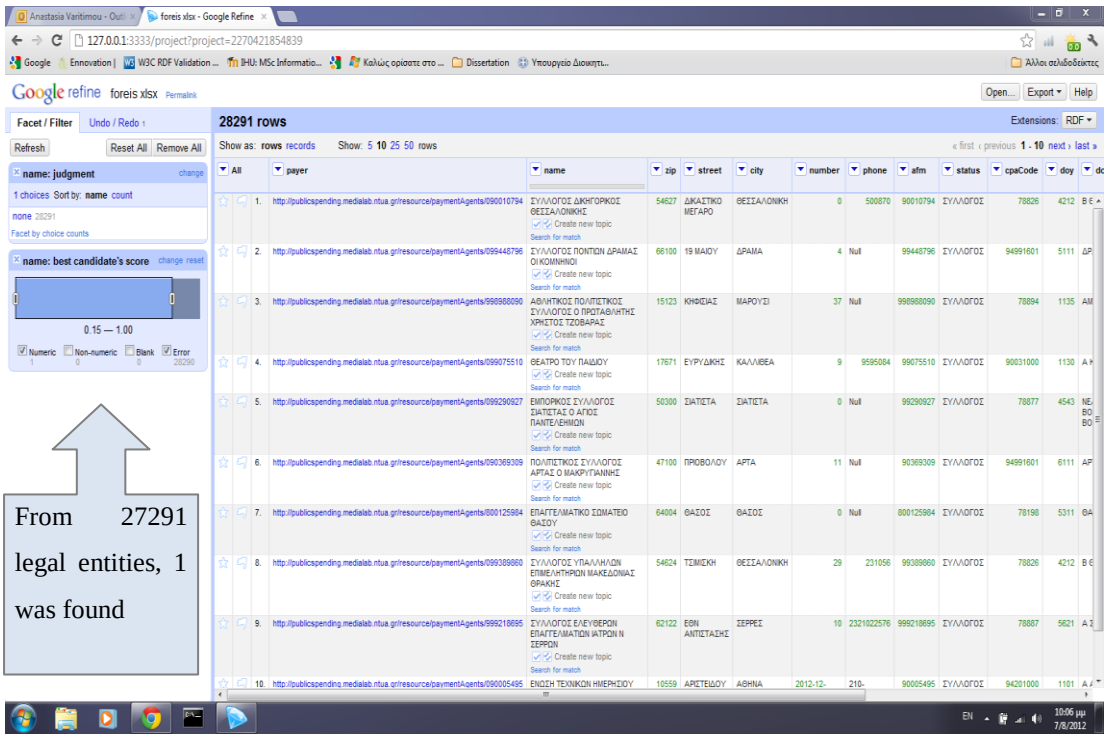


Figure 33 Reconcile names with Greek DBpedia ontology

Given the fact that Greek terms are rare in the Linked data cloud and pure results have been obtained from our search for *Greek* Linked Data with *Sindice* and in *CKAN* repository, we turned to geography. Geography is always an intuitive way to align datasets, especially, since geographical dimensions exists in almost every dataset and big hubs like Geonames[70] contain millions of URIs representing geospatial information from around the globe. We decided to use Geonames instead of the other geospatial Linked Data hubs, based on the translation capabilities its ontology offers.

Geonames does not provide a SPARQL endpoint. Instead it offers RDF web services free of charge up to a point (maxrows=1000 retrieved from the web service), through where we obtained the *RDF file describing all Greek features*. The format of the service was:

```
http://api.geonames.org/search?maxRows=1000&type=rdf&username=demo&country=gr
```

and we named the RDF file created, *GeonamesGreece.rdf*.

Geonames provides all geographical objects as *Features*. Each Feature is a specific Web recourse identified with a dereferenceable URI, and depending on its type belongs to a specific Feature class. Features belonging to *http://www.geonames.org/ ontology#P*[71]

[70] http://www.geonames.org/ [Accessed September 2012]

[71] Described in Geonames ontology http://www.geonames.org/ontology/ontology_v3.01.rdf

are cities, town or villages. For example the feature describing city of Thessaloniki goes as follows:

```
<gn:Feature rdf:about="http://sws.geonames.org/734077/">
  <rdfs:isDefinedBy>http://sws.geonames.org/734077/about.rdf</rdfs:isDefinedBy>
  <gn:name>Thessaloniki</gn:name>
  <gn:alternateName xml:lang="ar">سالونيك</gn:alternateName>
  <gn:alternateName xml:lang="bg">Солун</gn:alternateName>
  <gn:alternateName xml:lang="bs">Solun</gn:alternateName>
  <gn:alternateName xml:lang="ca">Tessalònica</gn:alternateName>
  <gn:alternateName xml:lang="cs">Soluň</gn:alternateName>
  <gn:alternateName xml:lang="cu">Селунь</gn:alternateName>
  <gn:alternateName xml:lang="de">Thessaloniki</gn:alternateName>
  <gn:alternateName xml:lang="el">Σαλονικη</gn:alternateName>
  <gn:alternateName xml:lang="en">Thessaloniki</gn:alternateName>
  <gn:alternateName xml:lang="es">Salónica</gn:alternateName>
  <gn:alternateName xml:lang="et">Thessaloníki</gn:alternateName>
  <gn:alternateName xml:lang="fi">Thessaloniki</gn:alternateName>
  <gn:alternateName xml:lang="fr">Thessalonique</gn:alternateName>
  <gn:alternateName xml:lang="he">סלוניקי</gn:alternateName>
  <gn:alternateName xml:lang="hr">Solun</gn:alternateName>
  <gn:alternateName xml:lang="id">Thessaloniki</gn:alternateName>
  <gn:alternateName xml:lang="is">Þessalóníka</gn:alternateName>
 >
  <gn:alternateName xml:lang="tr">Selânik</gn:alternateName>
  <gn:alternateName xml:lang="uk">Салоніки</gn:alternateName>
  <gn:alternateName xml:lang="zh">塞萨洛尼基</gn:alternateName>
  <gn:featureClass rdf:resource="http://www.geonames.org/ontology#P"/>
  <gn:featureCode rdf:resource="http://www.geonames.org/ontology#P.PPLA"/>
  <gn:countryCode>GR</gn:countryCode>
  <gn:population>354290</gn:population>
  <wgs84_pos:lat>40.64028</wgs84_pos:lat>
  <wgs84_pos:long>22.94389</wgs84_pos:long>
  <gn:parentCountry rdf:resource="http://sws.geonames.org/390903/"/>
  <gn:nearbyFeatures rdf:resource="http://sws.geonames.org/734077/nearby.rdf"/>
  <gn:locationMap
rdf:resource="http://www.geonames.org/734077/thessaloniki.html"/>
</gn:Feature>
```

The joint part between the RDF file from Geonames and our RDF describing companies was the concepts of city, village or key location where legal entities reside, described in our schema with the term *postName* of the Core location Vocabulary and with the term *alternateName* in Geonames.

5.3.5 Integration

Interlinking our data with related datasets from Publicspending and Geonames, discovered through the previous phase, will turn them to 5-star Linked Data.

The terms that will be used for the linking depends on the nature of the data and the relationship between them, while the whole process can be completed automatically or manually.

Interlinking with Publicspending's Linked Data

As stated in discovery phase, an obvious outgoing link would be to state that our URIs representing legal entities actually refer *to the same real-world concept* with URIs describing companies in Publicspending. To create URIs aliases like that, the *owl:sameAs* property is widely used. Creating such an identity link, was fairly easy, since we included it to the publishing procedure of our RDF file with Google Refine RDF extension, using the language GREL as it can be seen in the RDF description of an example company.

```
<rdf:Description
rdf:about="http://linkeddata.ihu.edu.gr/resource/company/90010794">
    <rdf:type rdf:resource="http://www.w3.org/ns/legal#legalEntity"/>
    <legal:legalName>ΣΥΛΛΟΓΟΣ ΔΙΚΗΓΟΡΙΚΟΣ ΘΕΣΣΑΛΟΝΙΚΗΣ</legal:legalName>
    <legal:companyStatus
rdf:resource="http://lInkeddata.ihu.edu.gr/resource/status/NormalActivity"/>
    <legal:companyType
rdf:resource="http://lInkeddata.ihu.edu.gr/resource/grtypes/assoc"/>
    <legal:legalIdentifier
rdf:resource="http://lInkeddata.ihu.edu.gr/resource/company/li90010794"/>
    <legal:registeredAddress
rdf:resource="http://lInkeddata.ihu.edu.gr/resource/company/ra90010794"/>
    <owl:sameAs
rdf:resource="http://publicspending.medialab.ntua.gr/resource/paymentAgents/90010
794"/>
  </rdf:Description>
```

Interlinking with Geonames Linked Data

Creating links between the locations of the companies in our dataset with locations from Geonames.org, has proved to be a more complex procedure that linking with Publicspending.

While integrating two datasets with common terms is easy, it becomes a little more complicated when similarity approaches should be taken into consideration. The lack of variety in Greek terms described as Linked Data can make the whole procedure even more difficult.

In our dataset, cities and key locations are described with uppercase Greek characters and no punctuation, while in Geonames dataset are provided with lowercase and with punctuation. Google Refine Reconciliation Services could not find many similarities mainly due to the punctuation difference. Figure 34 illustrates the result of the reconciliation services based on Sindice search in Geonames.org where only 508 records from a total of 28291 were reconciled.

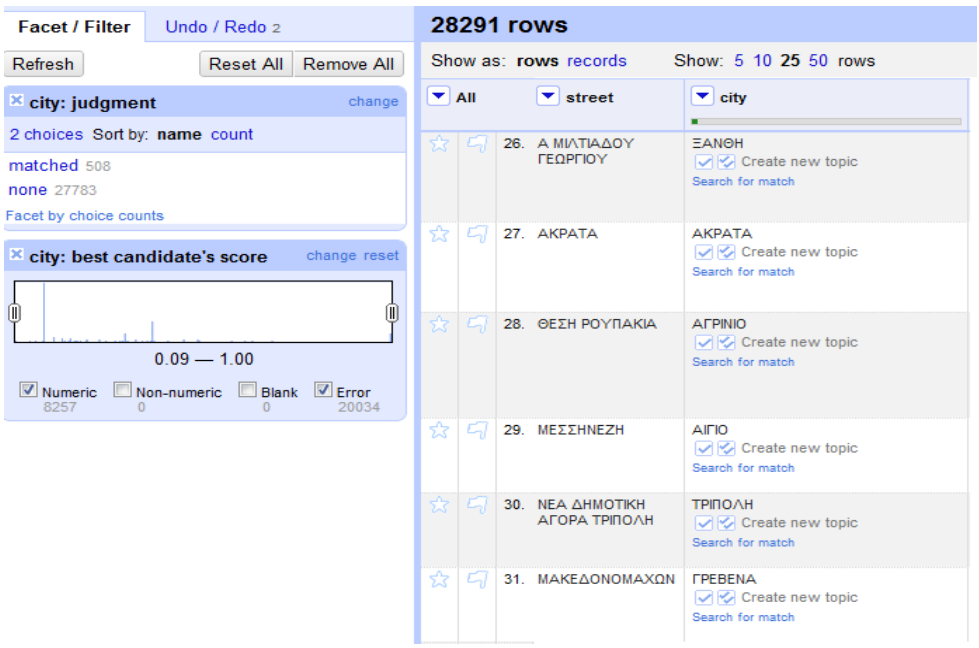


Figure 34 Reconciliation results with Sindice search in Geonames

As we concluded, Google Refine Reconciliation Services with Geonames didn't have a problem with the uppercase-lowercase difference between the datasets but could not overcome the punctuation difference. There were also other differences coming from wrong writing of Greek cities in Geonames. Table 10 describes some of these differences between the joint terms of the two datasets.

Our new Linked Data ***Companies.rdf***	*Geonames Linked Data* ***GeonamesGreece.rdf***	***Differences***
ΒΟΛΟΣ	Βόλος	Lowercase, punctuation
ΑΘΗΝΑ	Αθήναι	Lowercase, punctuation, old writing
ΘΕΣΣΑΛΟΝΙΚΗ	Σαλονικη	Lowercase, punctuation, wrong writing

Table 10 Differences describing cities in our RDF file and Geonames

So, our resources could not be easily interlinked with resources from Geonames because the joint terms are *similar* and not *identical*.

SILK is a framework especially designed to provide similarity-approach linkage heuristics based on transformations and multiple comparisons between different properties of two entities. The results are aggregated and if the resulted score is above a threshold, the two entities are interlinked.

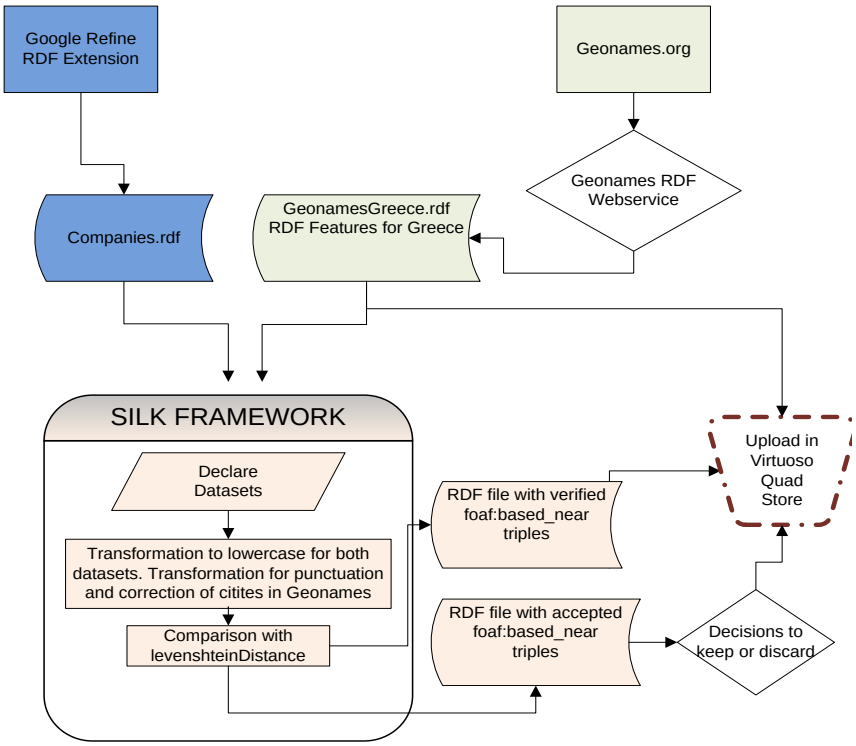


Figure 35 The process of interlinking with SILK

With SILK we overcame all the obstacles created by the different formats of describing cities, villages and key locations between the Geonames file and our file. Figure 35 illustrates the procedure followed.

SILK uses a special language called *Silk Link Specification Language (SILK-LSL).* With SILK-LSL you first have to define the datasets that you want to interlink. The data can reside behind a SPARQL endpoint or can be stored in a RDF file. We used the RDF file that we had obtained through Geonames.org *(GeonamesGreece.rdf)* and our RDF file describing companies we had create from Google Refine. The declaration of the datasets using SILK-LSL was as follows:

```
<DataSources>
    <DataSource id="myCity" type="file">
      <Param name="file" value="Companies.rdf" />
      <Param name="format" value="RDF/XML" />
    </DataSource>

    <DataSource id="geoNameCity" type="file">
      <Param name="file" value="geonamesGreece.rdf" />
      <Param name="format" value="RDF/XML" />
    </DataSource>
  </DataSources>
```

The specific resources that will be interlinked are provided through restriction rules on the datasets. We declared that we wanted to create links between the *Addresses* of our

companies with the resources of type *Features* in Geonames as described in the code below:

```
<SourceDataset dataSource="myCity" var="a">
      <RestrictTo>
        ?a rdf:type locn:Address
      </RestrictTo>
    </SourceDataset>

    <TargetDataset dataSource="geoNameCity" var="b">
      <RestrictTo>
        ?b rdf:type gn:Feature
      </RestrictTo>
    </TargetDataset>
```

Defining the type of link that should be created between the two datasets was in order. The more often used link is *owl:sameAs* but this wasn't our case. We wanted to state that our companies are *based in* locations described in Geonames so we used *foaf:based_near* which is a common term from FOAF[72] vocabulary to relate two spatial things.

For the links to be created between resources, SILK-LSL needs to know the exact term for the comparison and based on what algorithm will these terms be compared. If needed, transformations may occur to the items before the comparison. We had to performed extended transformations to eliminate the problems of uppercase, punctuations and old-writing to be able to compare the selected terms using *the levensteinDistance* algorithm defining as threshold =1. According to this algorithm and the threshold we have set, the resources would be interlinked if there was only up to a character difference between compared items. The comparison and some of the transformation parameters were declared with SILK-LSL as follows:

```
<Compare metric="levenshteinDistance" threshold="1">

                <TransformInput function="lowerCase">
                  <Input path="?a/locn:postName" />
                </TransformInput>

                  <TransformInput function="replace">
                       <TransformInput function="replace">
                          <Input path="?b/gn:alternateName[@lang='el']" />
                          <Param name="search" value="Αθήναι" />
                          <Param name="replace" value="αθηνα" />
                        </TransformInput>
                        <Param name="search" value="Σαλονικη" />
                        <Param name="replace" value="θεσσαλονικη" />
                   </TransformInput>
   </Compare>
```

[72] Foaf vocabulary specifications http://xmlns.com/foaf/spec/ [Accessed September 2012]

Our final output was two RDF files in RDF/N3 syntax containing a total of 21.272 *foaf:based_near* triples that indicate that our companies are based near locations described in Geonames, like the ones listed below:

```
<http://linkeddata.ihu.edu.gr/resource/company/ra94033829>
<http://xmlns.com/foaf/0.1/based_near>  <http://sws.geonames.org/264371/> .
<http://linkeddata.ihu.edu.gr/resource/company/ra999841409>
<http://xmlns.com/foaf/0.1/based_near>  <http://sws.geonames.org/7522530/> .
<http://linkeddata.ihu.edu.gr/resource/company/ra999637970>
<http://xmlns.com/foaf/0.1/based_near>  <http://sws.geonames.org/736083/> .
```

We preferred to create two output files instead of one to distinguish between triples with a high level of certainty (*Geonames_cities_verified.rdf* with 16465 triples *)* and triples that should be examined before uploading them to Virtuoso (*Geonames_cities_accepted.rdf* with 4807 triples).

Finally our two output RDF files together with the RDF file obtained by Geonames.org were uploaded to Virtuoso and shared in our package in CKAN.

Figure 36 gives an example of the links that were created during the phase of integration for a single company. Interlinking our companies with Geonames provides us with further knowledge about our legal entities. Linked data clients or browsers that search for a company or a public body can select its address and by following the link *foaf:based_near* obtain further informations for the city that is based near as depicted in Figure 36. Visualizations can be created based on the geographical coordinates now available. They can even move up in the geographical hierarchy from a city to the country and create mashups to the country as a whole. The reverse procedure can also be followed; by searching a city, someone can find the legal entities that reside there and informations about them. Additional informations about payments and decisions can also be obtained by following the link to Publicspending's resources.

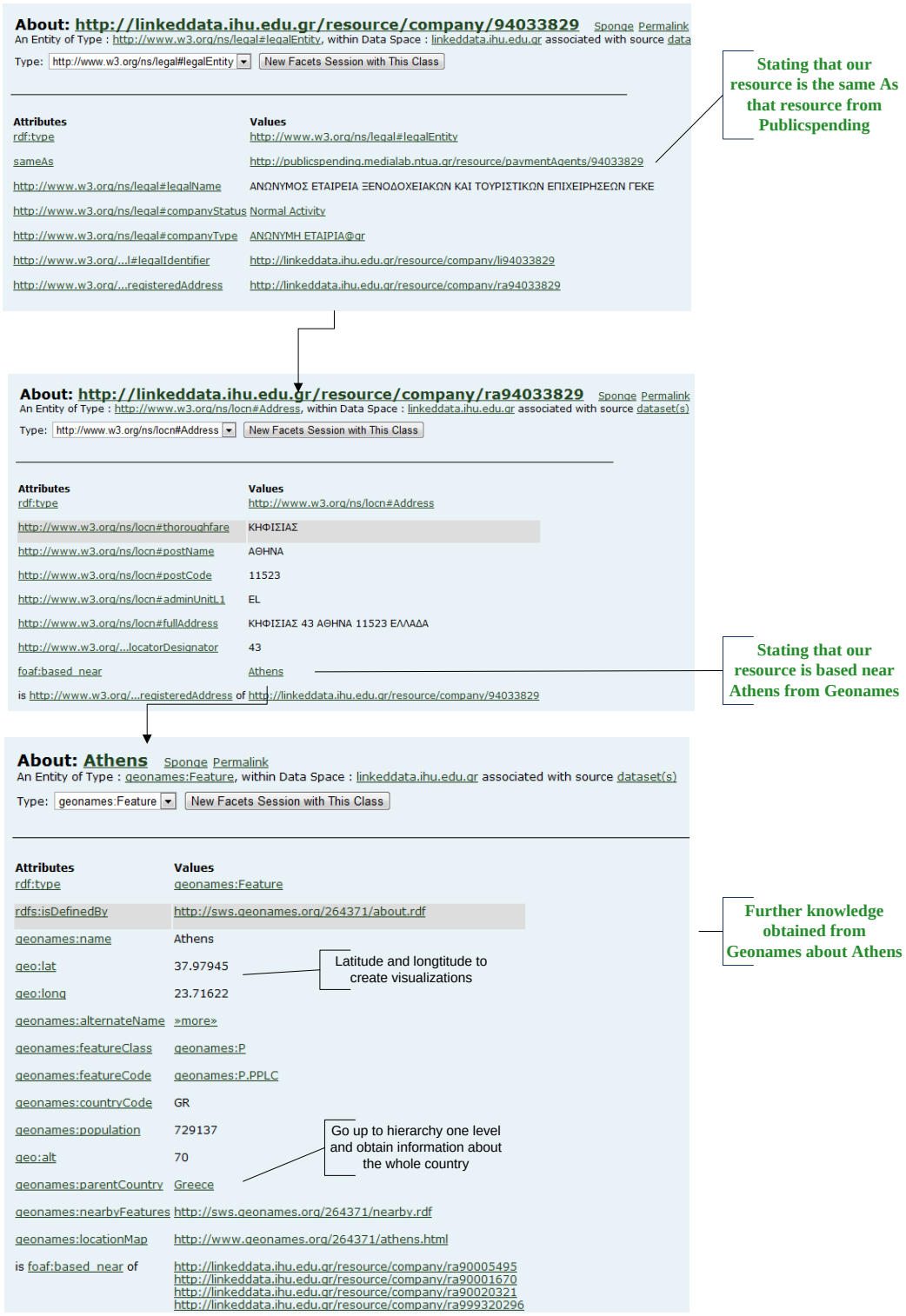


Figure 36 Links with Publicspending and Geonames

The advantages do not end here. Geonames is a big geographical hub in the Linked Data cloud. Its features have been interlinked with other Linked Data hubs, like DBpedia, LinkedMDB, New York Times etc. Linked Data browsers can follow these links and discover more information. In Figure 37 we can see the results that the *Tabulator browser* has produced when we searched for the specific Geonames feature representing

Athens. Without having implicitly created links between our resources and DBpedia, we can gain further knowledge through the mechanisms of the Web of Data.

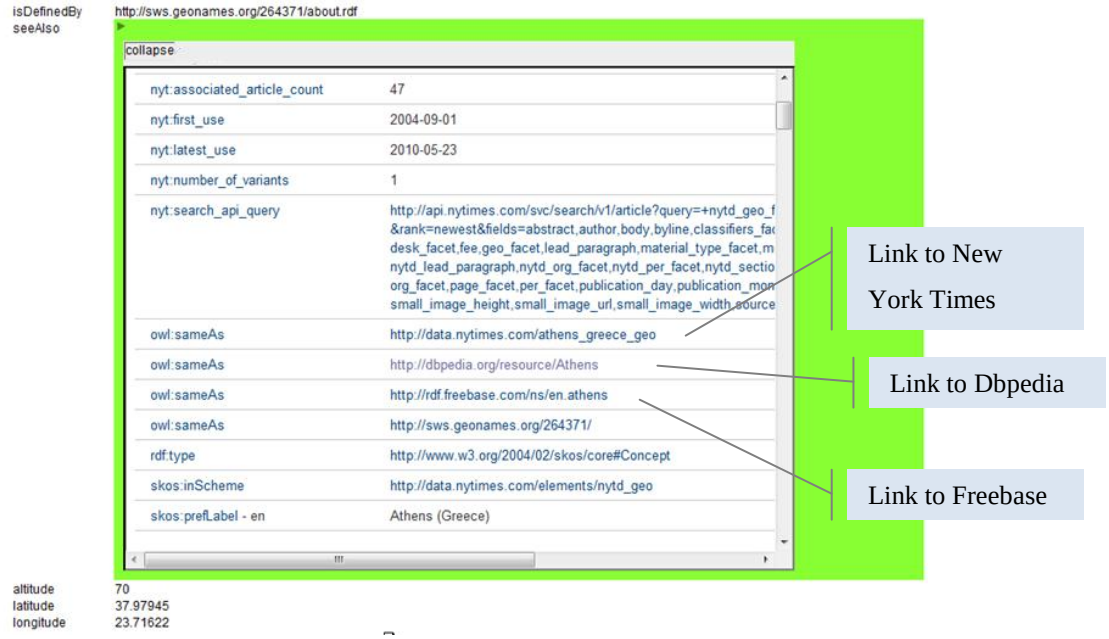


Figure 37 Tabulator results for Geonames Feature "Athens"

5.3.6 Data Browsing

We have until now completed the publishing of our initial CSV dataset as Linked data and further more we turned it to 5-star Linked data by integrating it into the linked data cloud. Linked Data browsers and search engines can now navigate and discover our data by following RDF links, just as traditional web browsers follow HTML links. Unlimited applications can be developed based not only from the standard format we used to describe legal entities but also, from the additional information that can be exploited through links.

Since not all of these applications can be created within this thesis, we decided to demonstrate the use of our data by deploying a faceted browser to present our data to non-experts end-users. However, many other useful Linked Data sites can also be developed to consume informations provided through our sparql endpoint, by combining informations from the interlinked datasets and visualizations can be created due to the geographical dimension of our data.

Faceted browsers provide an intuitive interface to present data that differs from the traditional text-match searching and browsing. It aggregates results based on facets and presents to the user only the data that match the attributes of a particular facet. Through a human-friendly interface, users are guided to create queries to data, stored in RDF files or SPARQL endpoints, without the need to be an expert in the field. Faceted

browsers work on top of these data only and cannot navigate through links like a Linked Data browser can but saves the end-users from the chaotic details of RDF and SPARQL.

Instead of creating yet another faceted browser we decided to use an open source faceted browser developed in DERI that works on top of SPARQL endpoints and provides a Google Refine-like interface. The browser can easily be configured by editing two JSON files. The first of these files is used to declare the facets that the user would like to use while in the other the presented items are defined. The browser didn't support Unicode characters and since our data are based solely in Greek characters, considerable changes in the code had to be made.

As facets we used the company type and the city that the company resides but since they are not included in the main triplet of the legal entity, we had to create more complex triples as defined in the JSON file:

```
[
    {
   "name":"Κατηγορίες εταιριών",
   "varname":"pages",
   "filter": {"pattern":"<http://www.w3.org/ns/legal#companyType> ?type.
?type <http://www.w3.org/2004/02/skos/core#notation>"}
  },
    {
   "name":"Πόλεις που ανήκουν",
   "varname":"city",
   "filter": {"pattern":"<http://www.w3.org/ns/legal#registeredAddress> ?addr
. ?addr <http://www.w3.org/ns/locn#postName>"}
    }
]
```

The best part of the browser is that offers the ability to the user, from the client side, to add his own facets, thus create his queries and becomes completely parameterized in the client side.

For the browser to be deployed a java application server was in need and for that purpose we select *Glassfish*. Figure 38 illustrates our browser's results when only the companies with company type ΝΠΔΔ and the city ΘΕΣΣΑΛΟΝΙΚΗ were requested through the first and second facet respectively.

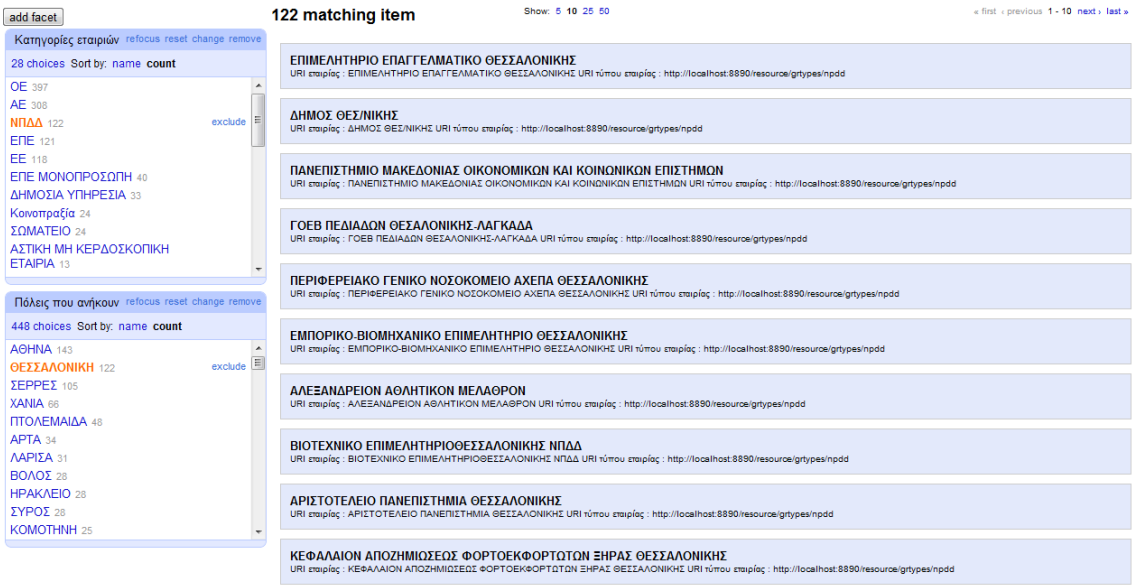


Figure 38 Greek Legal Entities Faceted Browser

If a user wanted to see the same results without the use of our faceted browser, he would have to create the following query to our sparql endpoint:

```
select ?s ?name ?skostype where {?s a <http://www.w3.org/ns/legal#legalEntity>;
<http://www.w3.org/ns/legal#legalName> ?name;
<http://www.w3.org/ns/legal#companyType> ?type.
?type <http://www.w3.org/2004/02/skos/core#notation> ?skostype.
FILTER(?skostype=str("ΝΠΔΔ")).
?s <http://www.w3.org/ns/legal#registeredAddress> ?address.
?address <http://www.w3.org/ns/locn#postName> ?city.
FILTER(?city=str("ΘΕΣΣΑΛΟΝΙΚΗ"))} ORDER BY ?s
```

As easily concluded, an end-user could greatly benefit from the deployment of a faceted browser since it allows the display of companies based on different queries without any expertise in SPARQL and in the modeling of our schema. Our browser also supports multiple facet combinations between cities and company types. Figure 39 illustrates all the companies that are public entities, thus have company type ΝΠΔΔ and reside in cities ΗΡΑΚΛΕΙΟ and ΒΟΛΟΣ

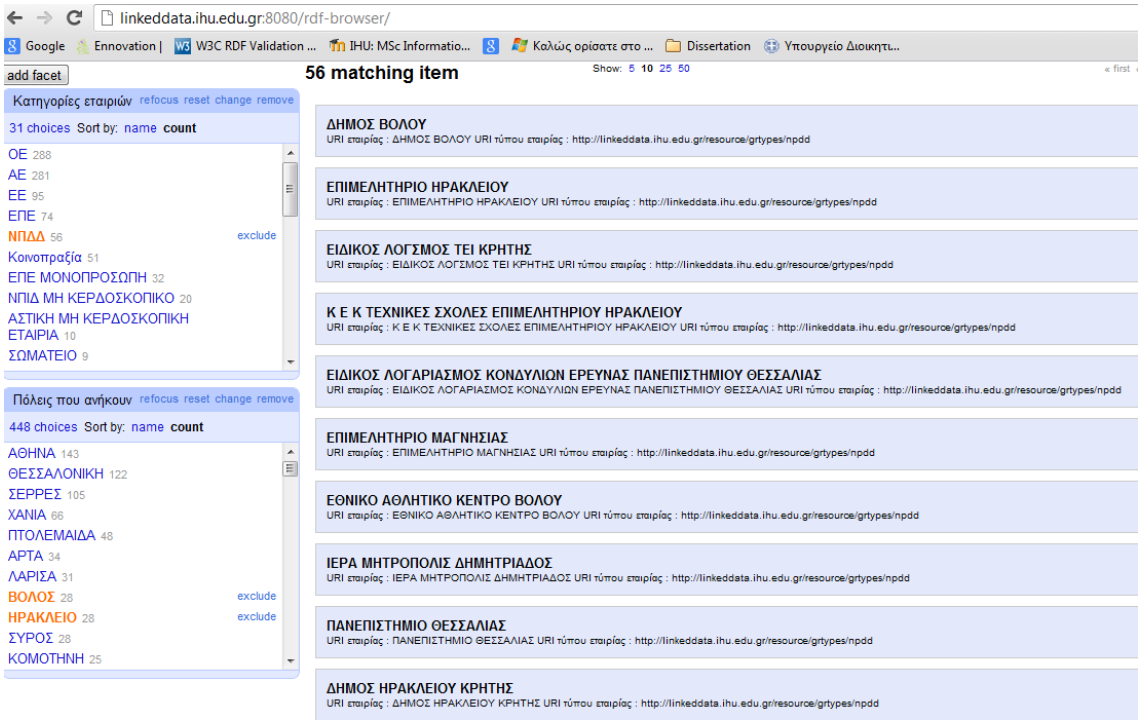


Figure 39 Multiple facet combinations in RDF Faceted Browser

If we didn't have the faceted browser, a user in order to retrieve the same informations would have to write the SPARQL query that follows:

```
select ?s ?name ?city ?skostype where {?s a <http://www.w3.org/ns/legal#legalEntity>;
<http://www.w3.org/ns/legal#legalName> ?name;
<http://www.w3.org/ns/legal#registeredAddress> ?address.
?address <http://www.w3.org/ns/locn#postName> ?city. FILTER(?city=str("ΗΡΑΚΛΕΙΟ") ||
?city=str("ΒΟΛΟΣ" )).
?s <http://www.w3.org/ns/legal#companyType> ?type. ?type
<http://www.w3.org/2004/02/skos/core#notation> ?skostype FILTER(?skostype=str("ΝΠΔΔ"))
} order by ?city
```

Our Faceted Browser is also hosted in International Hellenic University and can be accessed through the URL : http://linkeddata.ihu.edu.gr:8080/rdf-browser/

5.4 Publishing and Consumption pipeline overview

In this chapter, we described a pipeline for publishing of Greek OGD as RDF based on EU standards and its consumption through a Faceted Browser.

We initiated our pipeline by searching for available Greek OGd and we found our desired dataset in Publicspending's endpoint. The dataset described 28.892 Greek legal

entities and was in CSV format. Extended transformations using Excel functions were needed to create the desired format.

We decided on which vocabularies we would use to model our schema, based on the fact that we wanted our RDF triples to be interoperable and highly-reusable. So we used the **Core Business and Core Location Vocabularies** and we created a new controlled vocabulary called CompanyTypes.rdf to declare classes and schemas for company types and status using the SKOS concept.

To convert the CSV file to RDF we used Google Refine RDF Extension and we produced an RDF file, Companies.rdf containing 17 triples describing each of the 28892 Greek Legal Entities. To conclude the publishing phase, we uploaded the RDF files to Virtuoso's Quad Store, place them behind a SPARQL endpoint and used Virtuoso's URL-rewriter to make them dereferenceable.

To integrate our RDF data into the linked data cloud related datasets had to be discovered. Publicspending was an obvious candidate for interlinking since we both describe the same legal entities using different vocabularies. We also based our interlinking to the geography dimension; we obtained an RDF file from Geonames.org RDF API describing Greek features (Greekgeonames.rdf).

To create *owl:sameAs* links with Publicspending we included it in the RDF skeleton of the initial creation of our Companies.rdf file with Google Refine RDF Extension. To create links with Geonames, we had to use SILK framework due to considerable differences to the joint points between our datasets. We based our comparison of the two datasets on the *Levenshtein Distance Algorithm* and produced 2 files (Geonames_cities_accepted.xml and Geonames_cities_verified.xml) containing 21000 triples using as predicate the *foaf:based_near* term.

All our RDF files were uploaded in CKAN along with a new VoID file void.ttl which contains metadata about our datasets in a new package http://datahub.io/dataset/greek-legal-entities to enable discovery and consuming in applications.

To reveal the usefulness of our linked data we deployed a faceted browser. With this browser, even inexperienced users can browser our described with RDF legal entities and filter them based on company types, cities they resides or combination of both.

6 Conclusions and lessons learned

In this thesis we tackled the problem of data integration in developing OGD applications which is raised due to the different format of the data, the lack of sufficient metadata, the different semantics and the lack of common national and organizational characteristics.

In this direction we conducted a survey to gather the applications that have been developed worldwide to consume OGD, trying to find out the way the developers are facing these problems but also to identify trends and opportunities. Through this survey, 350 applications and metadata from their catalogues were yielded and among other useful conclusions, we identified that data integration was not apparent in most OGD apps.

We argue that applying Linked Data principles *with the use of wide-accepted and reusable vocabularies* on OGD can address this problem by providing data heterogeneity, a standardize manner in publishing OGD data, and facilitate consensus and semantic interoperability. It can then enable the integration of different datasets with final goal to motivate the development of more applications that will add value to the open data initiatives.

To support our thesis, we presented a publishing and consumption pipeline that transformed Greek Government Data to Linked Data using the proposed by the EU Core Vocabularies, we integrated it into the linked data cloud and we deployed a faceted browser to make it available to the end-users.

Describing legal entities with the Core Vocabularies provided semantics and enhanced their pan-European interpretation. It provided scalability and usability even for the citizens of the same country as it substitutes every plain literal, meaningless for non-experts in the field, with URIs that are in their turn described in a well-structured but also flexible way.

Furthermore, it enabled the creation of joint points with other external datasets and facilitated their interlinking with the linked open cloud especially with the use of the Identifier class and the description of the registered address with the Core Location Vocabu-

lary. Their advantages can further be recognized as more information about legal entities will be transformed to Linked Data.

Our work led us also to other conclusions, some anticipated and some more unexpected and posed future questions and work in a variety of fields:

— *OGD should be provided through direct access to central catalogues in raw formats.*

From our experience through this thesis, the most difficult process in our pipeline was the phase of discovery, since the absence of Greek OGD in structured formats gathered in the one single point, hindered any attempt of finding new inputs and designing an application on it. Even when data was available from web services, the different access mechanisms and the time acquired for retrieval was prohibiting.

Data that resides in scattered catalogues and websites require exhaustive browser search and its discovery can prove challenging. Even more discouraging is its discovery in non-structured formats. Mixing data with their presentations, like in HTML or PDF format, have use only for people but it is not machine readable. The information although *available* it is *not reusable*.

Our survey in the OGD world showed us, that where a single access point existed to provide all the available government datasets not only facilitated their discovery, but also enabled statistics, evaluation and better services.

— *The need for Linked Data is apparent.*

The full potential of Linked Data will be uncovered as more and more data is being published that will enable interlinking, further knowledge and integration. The need for an authoritative source that will serve as reference hub at least for Greek data was unveiled through our pipeline. Even more useful would be the creation of international or pan-European dataset hub with accurate geospatial information or other common features. At least, updating big hub's datasets with more accurate and up-to-date information like DBpedia Greek version and Geonames Greek dataset and enriching it in particular with postal codes should be a first priority since it is technically feasible and with immediate results.

— *Linked data technologies seem mature but workload still remains heavy. It requires considerably added effort for data publishers and end-user application developers.*

The technical community has created a number of powerful software tools that can support Linked Data publishing phases. Integration of phases in one single workbench like Google Refine RDF extension, try to ease the developer's workload and should be further encouraged. However the effort still remains heavy mainly due to 2 reasons:

First, the phase of modeling proved to be a challenging task, even more than data modeling in traditional information systems, as it required detailed analysis down to the data level and extended transformations that should be made upfront.

Second, all the phases for publishing and consuming Linked Data are added to the existing Web development processes. Even when Linked Data is already published, consuming them in end-user applications equals additional work. Maybe that is the main reason that the application developers community have not embrace Linked Data, look with suspicion to RDF and SPARQL and prefer immediate access to simple formats. Better tools and automation on consuming Linked Data, metadata catalogues and descriptions around SPARQL endpoints or RDF dump files should be taken under consideration.

Our research in the OGD application world gave us a holistic view on how OGD is being consumed around the globe. By exploring existing apps in detail, we collected information related to the most popular types of apps, their business model, the lifecycle of an app and programming languages used whenever possible. Almost all applications were accessed through catalogues and we identified the metadata model of these catalogues. Certain important factors were identified:

— *A common metadata model for application catalogues in those portals would facilitate their discovery.*

Metadata standards for describing applications are currently missing. All app catalogues seem to share some common features about their apps like: app name, URL, description and app category, while the other characteristics vary. Different classification regarding app category is also apparent although we were able to recognize 13 common major categories described in section 3.1.2. References to the datasets used, a powerful feature was characteristically missing from almost every catalogue.

Deciding on keeping common metadata model, e.g. by extending RADion, will facilitate the applications discovery and usage and give useful information inputs to the authorities. Features like the dataset that have been used, the version of the dataset and the last update date should be included in the model to improve trust and usability.

— *Many applications actually promote transparency, collaboration between citizens and authorities and improved government services, some of the main goals of OGD movement. However the business and commercial sectors don't appear to have embraced the Open Government Data app world.*

Through our survey in the OGD apps world, we identified excellent examples of applications that actually promote transparency and citizens' collaboration and could serve as powerful tools to whomever government wants to enhance its democratic procedures. However the amount of applications that are engaged in the commercial and businesses sector was exceptionally low. Is it the discovery and integration of relevant datasets across international catalogues that will enhance their interest and how can they be aware of the opportunities? Data aggregations and integrations on pan-European or even global level, presented through user-friendly interfaces could support businesses decision making, unveil opportunities, enhance firms' visibility, produce revenues and economic growth.

— *Central government catalogues that provide Open Government Data, seems also to be the perfect place for maintaining a catalogue with applications that were built on them.*

Concentrating datasets, applications and developers community interest, creates powerful ecosystems.

Finally, the fact that our schema was the first full implementation of the Core Business Vocabulary, identified problems that could only be unveiled in practice and defined a roadmap for future work.

— The concept of 66 different company types had to be declared in a new controlled vocabulary; in the rare cases that this was possible a translation with common company types was included to increase interoperability. Further work in this field should be done, easy for a technical point of view but far more complicated in the business aspect, so that *every Greek company type is mapped with a common EU schema.*

— Company status property is described in our schema but was not included in the original dataset; we assumed normal activity since the dataset was originally derived from General Secretariat of Information Systems of Ministry of Finance's web services. In cases where an end date existed, we considered the legal entities to be inactive. *Clarifying the possible values of the property* but also the ver-

sioning of the dataset is in our future plans since maintaining accurate up-to-date information is crucial for the re-use and trustworthiness of our data.

— Company activity is not included at present in our schema. Future work for the description of companies' activity should be investigated since even when we obtained the CPA code for every company, considering the fact that we had 28800 companies, the NACE codes for every section and subsection were impossible, at least at the boundaries of this thesis to be described in a new vocabulary. Applying Linked Data principles to the NACE coding system could ease the whole procedure since it would shift the problem to interlinking.

Despite the challenges, our work gave us very positive outcomes, significant feedback and proved our original statement: the transformation of Open Government Data to Linked Open Government data coupled with the use of standard and flexible vocabularies, provides data heterogeneity, semantic interoperability and addresses problems created due to domain and political boundaries. It facilitates data integration and opens the road for the creation of new innovative applications towards the goals of the Open Government movement.

Bibliography

[1] Advisory Board on Public Sector Information (2006*). Realizing the Value of Public Sector Information*. Retrieved from: http://www.nationalarchives.gov.uk/ documents/appsi-annual-report-2006.pdf. [Accessed May 2012].

[2] Australian Government, Office of the Australian Information Commissioner (2011). Issues Paper 2: *Understanding the value of public sector information in Australia.* Retrieved from: http://www.oaic.gov.au/publications/papers/issues_paper2_understanding_ value_public_sector_information_in_australia.html. [Accessed May 2012].

[3] The Memorandum on Transparency and Open Government. Retrieved from http://www.whitehouse.gov/the_press_office/TransparencyandOpenGovernment [Accessed June 2012]

[4] Open Government Partnership. Retrieved from http://www.opengovpartnership.org/ [Accessed June 2012]

[5] Bauer, F., Kaltenböck, M., 2011. *Linked Open Data: The Essentials.* Edition mono/monochrom, Vienna

[6] R. Pollock. *The Value of the Public Domain,* July 2006. Published by the Institute for Public Policy Research as part of a series on IP and the Public Sphere.

[7] M. Dekkers, F. Polman, R. te Velde, and M. de Vries. *MEPSIR: Measuring European Public Sector Information Resources.* 06/2006 2006

[8] Li Ding, Vassilios Peristeras, Michael Hausenblas, *"Linked Open Government Data,"* IEEE Intelligent Systems, pp. 11-15, May-June, 2012

[9] European Commission (2010). Digital Agenda for Europe 2010 - 2020 Retrieved from: http://ec.europa.eu/information_society/digital-agenda/index_en.htm. [Accessed June 2012]

[10] European Commission, DG INFSO (2010), eGovernment Action Plan 2011 – 2015 Retrieved from: http://ec.europa.eu/information_society/activities/egovernment/action_plan_2011_2015/index_en.htm. [Accessed June 2012]

[11] European Commission Public Sector Information (PSI) Directive. Retrieved from http://ec.europa.eu/information_society/policy/psi/rules/index_en.htm [Accessed June 2012]

[12] European Commission (2008). Results of the online consultation of stakeholders 'Review of the PSI directive'. Retrieved from: http://ec.europa.eu/information_ society/policy/psi/docs/pdfs/online_consultation/report_psi_online_consultaion_ stakeholders.pdf. [Accessed June 2012]

[13] European Commission 2011 Open Data Package Retrieved from:

http://ec.europa.eu/information_society/policy/psi/rules/index_en.htm [Accessed June 2012]

[14] GREEK ACTION PLAN 2012. Retrieved from Open Government Partnership http://www.opengovpartnership.org/countries/greece [Accessed July 2012]

[15] Tim Berners-Lee. Putting government data online, 2009. Retrieved from:

http:// www.w3.org/DesignIssues/GovData.html [Accessed June 2012]

[16] Tim Berners-Lee. Linked Data - Design Issues, 2006. Retrieved from: http://www.w3.org/DesignIssues/LinkedData.html [Accessed June 2012]

[17] Linking Open Data Project. Retrieved from http://www.w3.org/ wiki/ SweoIG/ Task-Forces/ CommunityProjects/LinkingOpenData [Accessed June 2012]

[18] CKAN – Open registry of data and content packages. Retrieved from http://datahub.io/ [Accessed October 2012}

[19] A. Bizer, C. Jentzsch and R. Cyganiak. *The Linking Open Data Cloud Diagram Webpage.* Retrieved from http://www4.wiwiss.fu-berlin.de/lodcloud/state/ [Accessed July 2012]

[20] UK Open Government Catalogue Website. Retrieved from http://www.data.gov.uk [Accessed July 2012]

[21] Shadbolt, Nigel, O'Hara, Kieron, Berners-Lee, Tim, Gibbins, Nicholas, Glaser, Hugh, Hall, Wendy and schraefel, m.c. (2012) *Open Government Data and the Linked Data Web: Lessons from data.gov.u*k. IEEE Intelligent Systems, Vol 27, Issue 3, Spring Issue.

[22] Alani, H.; Dupplaw, D.; Sheridan, J.; O'Hara, K.; Darlington, J.; Shadbolt, N.; and Tullo, C. 2007. *Unlocking the potential of public sector information with semantic web technology*. In ISWC/ASWC, 708–721.

[23] U.S Government Open Data Catalogue. Retrieved from http://www.data.gov [Accessed June 2012]

[24] Li Ding, Dominic DiFranzo, Sarah Magidson, Deborah L. McGuinness, and Jim Hendler - *Data-Gov Wiki: Towards Linked Government Data* ;In AAAI Spring Symposium: Linked Data Meets Artificial Intelligence(2010)

[25]: L. Ding, T. Lebo, J.S. Erickson, D. DiFranzo, G.T. Williams, X. Li, J. Michaelis, A.Graves, J.G. Zheng, Z. Shangguan, J. Flores, D.L. McGuinness, J. Hendler, - *TWC LOGD*: *A Portal for Linked Open Government Data Ecosystems*, Web Semantics:, vol. 9, no. 3, 2011, pp. 325-333.

[26] L. Ding, D. DiFranzo, A. Graves, J. Michaelis, X. Li, D. L. McGuinnessJ. A. Hendler. *TWC data-gov Corpus: Incrementally Generating Linked Government Data from data.gov. Proceeding WWW '10 Proceedings of the 19th international conference on World wide web Pages 1383-1386 2010*

[27] Government Linked Data Working Group. Retrieved from http://www.w3.org/2011/gld/ wiki/Main_Page [Accessed June 2012]

[28] European Commission FP7 ICT Program Retrieved from http://cordis.europa.eu/fp7/people/home_en.html [Accessed September 2012]

[29] LOD2. Retrieved from http://lod2.eu/Welcome.html. [Accessed June 2012]

[30] LATC (Linked Open Data Around-The-Clock). Retrieved from http://latc-project.eu/ [Accessed June 2012]

[31] LGD Life Cycle by Hausenblas. Retrieved from http://www.w3.org/2011/gld/wiki [Accessed June 2012]

[32] Li Ding Tim Finin Anupam Joshi Yun Peng R. Scott Cost Joel Sachs Rong Pan Pavan Reddivari Vishal Doshi *Swoogle: A Semantic Web Search and Metadata Engine* CIKM '04 Proceedings of the thirteenth ACM international conference on Information and knowledge management, Pages 652 - 659 2004

[33] Li Ding et al., *"Swoogle: A Search and Metadata Engine for the Semantic Web"*, Proceedings of the Thirteenth ACM Conference on Information and Knowledge Management, November 2004

[34] Cosmin Basca, Stéphane Corlosquet, Richard Cyganiak, Sergio Fernández, Thomas Schandl "*Neologism: Easy Vocabulary Publishing*", Proceedings of the Workshop on Scripting for the Semantic Web, in conjunction with ESWC 2008, 2008..

[35] Fadi Maali, Richard Cyganiak, Vassilios Peristeras - *A Publishing Pipeline for Linked Government Data* ESWC'12 Proceedings of the 9th international conference on The Semantic Web: research and applications, pages 778-792 Springer-Verlag Berlin, Heidelberg ©2012

[36] Christian Bizer, Freie Universität Berlin -*The Emerging Web of Linked Data* Intelligent Systems IEEE, 2009 Volume: 24 , Issue: 5 Page(s): 87 - 92

[37]T. Heath and C. Bizer, *"Linked Data: Evolving the Web into a Global Data Space,"* Synthesis Lectures on the Semantic Web: Theory and Technology, J. Hendler and Y. Ding, eds., Morgan & Claypool, 2011, pp. 1–136..

[38] Kurt Rohloff, Mike Dean, Ian Emmons, Dorene Ryder and John Sumner: *An Evaluation of Triple-Store Technologies for Large Data Stores* On the move to meaningful internet system OTM 2007 WORKSHOPS Lecture Notes in Computer Science, 2007, Volume 4806/2007, 1105-1114, DOI: 10.1007/978-3-540-76890-6_38

[39] *Benchmarking the Performance of Storage Systems that expose SPARQL Endpoints* Chris Bizer, Andreas Schultz Presented at: 4th International Workshop on Scalable Semantic Web Knowledge Base Systems, 2008

[40] Aidan Hogan, Andreas Harth, Jurgen Umbrich, Sheila Kinsella, Axel Polleres, Stefan Decker -*Searching and Browsing Linked Data with SWSE: the Semantic Web Search Engine* Web Semantics: Science, Services and Agents on the World Wide Web

[41] Christian Bizer and Andreas Schultz.- *The r2r framework: Publishing and discovering mappings on the web.* In Proceedings of the 1st International Workshop on Consuming Linked Data, 2010.

[42] Robert Isele, Anja Jentzsch, and Christian Bizer. -*Silk server - adding missing links while consuming linked data.* In Proceedings of the 1st International Workshop on Consuming Linked Data (COLD 2010), 2010

[43] TWC LOGD portal Retrieved from http://logd.tw.rpi.edu/ [Accessed September 2012]

[44] ISA_eGovernment-Core-Vocabularies_February2012.pdf. Retrieved from http://joinup.ec.europa.eu/category/highlight/national-interoperability-framework-observatory [Accessed July 2012]

[45] Core Vocabularies Business Location Person v1.00 Conceptual Model Retrieved from https://joinup.ec.europa.eu/asset/core_business/release/100 [Accessed August 2012]

[46] Core Person Vocabulary. Described in https://joinup.ec.europa.eu/asset/core_person/description [Accessed August 2012]

[47] Core Location Vocabulary. Described in https://joinup.ec.europa.eu/asset/core_location/description [Accessed August 2012]

[48] Core Business Vocabulary. Described in https://joinup.ec.europa.eu/asset/core_business/description [Accessed August 2012]

[49] Core Vocabularies Business Location Person v1.00 Specification. Retrieved from https://joinup.ec.europa.eu/asset/core_business/release/100 [Accessed August 2012]

[50] L. Sauermann and R. Cyganiak. *Cool URIs for the Semantic* Web.World Wide Web Consortium, Note NOTE-cooluris-20081203, December 2008.

[51] Alexander, K., Cyganiak, R., Hausenblas, M., Zhao, J. (2011). *Describing Linked Datasets with the VoID Vocabulary*. W3C Interest Group Note. Retrieved from: http://www.w3.org/TR/void/. [Accessed May 2012].

[52] Maali, F., Cyganiak, R., Peristeras, V. (2010). *Enabling Interoperability of Government Data Catalogues.* In Proceedings of the Electronic Government 10th International Conference, EGOV 2010, Lausanne, Switzerland

[53] Breitman, K. Salas, P. ; Casanova, M.A. ; Saraiva, D. ; Viterbo, J. ; Magalhães, R.P. ; Franzosi, E. ; Chaves, M. *Open government data in Brazil*, Intelligent Systems IEEE, May-June 2012, Vol. 26, Issue 3, p. 45-49

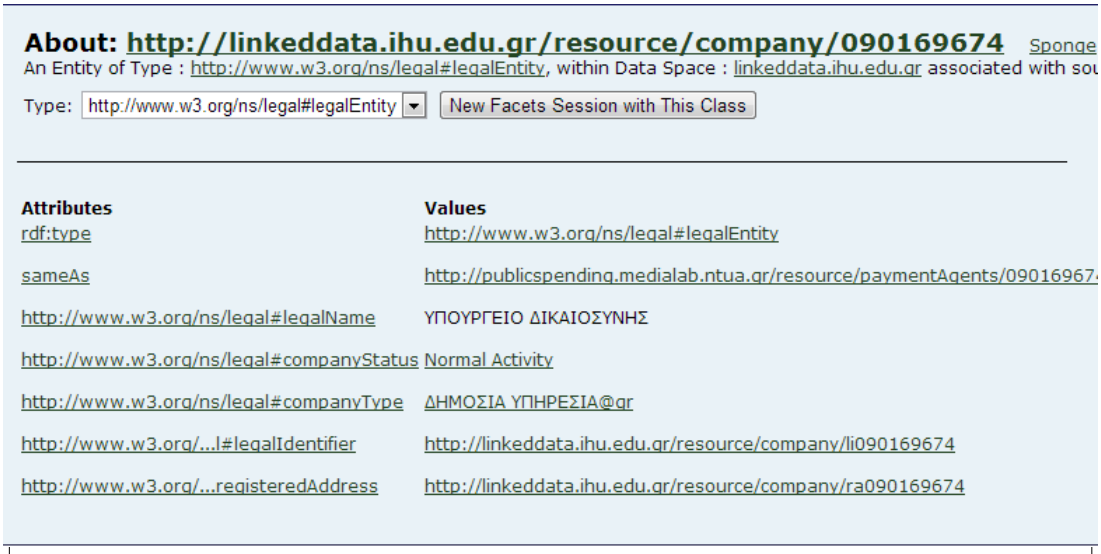
[54] Timothy Lebo, John S. Erickson, Li Ding, Alvaro Graves, Gregory Todd Williams, Dominic DiFranzo, Xian Li,James Michaelis, Jin Guang Zheng and Johanna Flores, *et al. Producing and Using Linked Open Government Data in the TWC LOGD Portal,* Linked Government Data *2011* Springer New York, Computer Science p.51-72

Appendix A

RDF for a Legal Entity

Example of our Linked Data representation of a single Legal Entity named ”ΥΠΟΥΡΓΕΙΟ ΔΙΚΑΙΟΣΥΝΗΣ” **modeled with the Core Vocabularies** (snapshots of the URI dereferenced and corresponding RDF code)

Legal Entity




```
<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/resource/company/090169674">
        <rdf:type rdf:resource="http://www.w3.org/ns/legal#legalEntity"/>
        <legal:legalName>ΥΠΟΥΡΓΕΙΟ ΔΙΚΑΙΟΣΥΝΗΣ</legal:legalName>
        <legal:companyStatus rdf:resource="http://linkeddata.ihu.edu.gr/resource/status/NormalActivity"/>
        <legal:companyType rdf:resource="http://linkeddata.ihu.edu.gr/resource/grtypes/dy"/>
        <legal:legalIdentifier rdf:resource="http://linkeddata.ihu.edu.gr/resource/company/li090169674"/>
        <legal:registeredAddress rdf:resource="http://linkeddata.ihu.edu.gr/resource/company/ra090169674"/>
        <owl:sameAs rdf:resource="http://publicspending.medialab.ntua.gr/resource/paymentAgents/090169674"/>
</rdf:Description>
```

Legal Identifier of the Legal Entity


```
<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/resource/company/li090169674">
        <rdf:type rdf:resource="http://www.w3.org/ns/adms#Identifier"/>
        <skos:notation>90169674</skos:notation>
        <adms:schemaAgency>Greek Tax System</adms:schemaAgency>
        <dc:created rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1988-01-10</dc:created>
        <dc:creator>ΔΟΥ ΙΒ ΑΘΗΝΩΝ</dc:creator>
</rdf:Description
```

Registered address of the Legal Entity


```
<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/resource/company/ra090169674">
        <rdf:type rdf:resource="http://www.w3.org/ns/locn#Address"/>
        <locn:thoroughfare>ΜΕΣΟΓΕΙΩΝ</locn:thoroughfare>
        <locn:locatorDesignator>96</locn:locatorDesignator>
        <locn:postName>ΑΘΗΝΑ</locn:postName>
        <locn:postCode>11527</locn:postCode>
        <locn:adminUnitL1>EL</locn:adminUnitL1>
        <locn:fullAddress>ΜΕΣΟΓΕΙΩΝ 96 ΑΘΗΝΑ 11527 ΕΛΛΑΔΑ</locn:fullAddress>
</rdf:Description>
```

Company status described in new controlled vocabulary


```
<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/def/code/status">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-      schema#Class"/>
<skos:prefLabel> Company activity</skos:prefLabel>
<rdfs:comment>This is the class of Greek company ypes</rdfs:comment>
<rdfs:subClassOf >http://www.w3.org/2004/02/skos/core#Concept</rdfs:subClassOf >
</rdf:Description>

<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/resource/status/NormalActivity">
            <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#ConceptScheme"/>
            <rdf:type rdf:resource="http://linkeddata.ihu.edu.gr/def/code/status"/>
<skos:prefLabel>Normal Activity</skos:prefLabel>
<skos:notation>Normal Activity</skos:notation>
</rdf:Description>
```

Company type described in new controlled vocabulary


```
<!--Declare the class and the concept schema for Company Type -->
<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/def/code/grtypes">
  <rdf:type rdf:resource="http://www.w3.org/2000/01/rdf-      schema#Class"/>
<skos:prefLabel> Greek company types</skos:prefLabel>
<rdfs:comment>This is the class of Greek company types</rdfs:comment>
<rdfs:subClassOf >http://www.w3.org/2004/02/skos/core#Concept</rdfs:subClassOf >
</rdf:Description>

<rdf:Description rdf:about="http://linkeddata.ihu.edu.gr/resource/grtypes/dy">
            <rdf:type rdf:resource="http://www.w3.org/2004/02/skos/core#ConceptScheme"/>
            <rdf:type rdf:resource="http://linkeddata.ihu.edu.gr/def/code/grtypes"/>
<skos:prefLabel>ΔΗΜΟΣΙΑ ΥΠΗΡΕΣΙΑ@gr</skos:prefLabel>
<skos:prefLabel>Public Service@en</skos:prefLabel>
<skos:notation>ΔΗΜΟΣΙΑ ΥΠΗΡΕΣΙΑ</skos:notation>
</rdf:Description>
```

Based near (reconciling city from registered address with Geonames)

About: Athens Sponge Permalink
An Entity of Type : geonames:Feature, within Data Space : linkeddata.ihu.edu.gr associated with source dataset(s)

Type: geonames:Feature | New Facets Session with This Class

Attributes	Values
rdf:type	geonames:Feature
rdfs:isDefinedBy	http://sws.geonames.org/264371/about.rdf
geonames:name	Athens
geo:lat	37.97945
geo:long	23.71622
geonames:alternateName	»more»
geonames:featureClass	geonames:P
geonames:featureCode	geonames:P.PPLC
geonames:countryCode	GR
geonames:population	729137
geo:alt	70
geonames:parentCountry	Greece
geonames:nearbyFeatures	http://sws.geonames.org/264371/nearby.rdf
geonames:locationMap	http://www.geonames.org/264371/athens.html
is foaf:based_near of	http://linkeddata.ihu.edu.gr/resource/company/ra90005495 http://linkeddata.ihu.edu.gr/resource/company/ra90001670 http://linkeddata.ihu.edu.gr/resource/company/ra90020321 http://linkeddata.ihu.edu.gr/resource/company/ra999320296 http://linkeddata.ihu.edu.gr/resource/company/ra90107045 »more»

```
<http://linkeddata.ihu.edu.gr/resource/company/ra90169674>  <http://xmlns.com/foaf/0.1/based_near>  <http://sws.geonames.org/264371/>
```

Appendix B

Open Data Apps survey

Sample of the Open Data Apps List yielded from our survey

A	B	C	D	E	F	G	H	I	J	L	O
Name	URL	Description	Publish	Catalogue	Programmir Language	Delivery mode (e.g. open source, fee)	Type (e.g. mobile, web, desktop)	Category	Data sets used	Usage info (e.g. download or likes)	Intendec audienc
100 People	http://data.gov.au/ap people/	Twelve point three million people sent their tax returns to the Australian Taxation Office (ATO) in the 2009 income year. In 100 people the ATO represents the millions of tax returns as a Flash animation of 100 people		data.gov.au	Flash	free to use	Web App	Business, Finance, Government, Industry, Society	Taxation Statistics 2007-08: Activity Statement Ratio Tables (Companies)	N/A	citizens of Australia
Know Where You Live	http://www.hackdays	Know Where You Live is the prototype of a data mashup that displays a range of Australian Government data based on your geographic location along with a Google satellite map and relevant photo from the Powerhouse Museum Collection, the State Records NSW or the State Library of New South Wales collection.	N/A	data.gov.au	Javascript / google maps	free to use	Web App	Community, Culture, General, Law, Safety, Society	NSW Crime Data, State Library of NSW Collection Photostream, State Records Collection Photostream (NSW), Subset of Powerhouse Museum Photographic Collection in Flickr Commons	N/A	citizens of Australia
Ottawa InsideOut	http://www.ottawa-events.ca/	Why not put all the event calendar data, museums, parks, sports and recreational information in one app? Surely a tourist visiting Ottawa or even residents of the city would like to have one application to encompass all the information they need about the city instead of downloading 4-5 apps. That is why I created this app. Google Maps, StreetView integration, the app works seamlessly and uses live event data published by the city	Samir	Ottawa open data apps directory	N/A	Android Market	Mobile	Art	Ottawa city open data	0 tweets + 0 fb likes	citizens of Ottawa
US GAAP RSS Feed of XBRL Financials	data.gov	This is an hourly update of the most recent Interactive Data documents submitted under the "Interactive Data to Improve Financial Reporting" rule (Release No. 33-9002) using US GAAP as the base taxonomy.	data.gov	data.gov / goverment apps	Rss Feed	free to use	Web App	Banking, Finance, and Insurance	"Interactive Data to Improve Financial Reporting" rule (Release No. 33-9002)	21534 Visits + 19535 downloads	citizens of USA
Can I Park It?	http://canipark.it		Eric George	data.gov.uk		free	Web App	Bikes	TFL Cycle Hire		citizens in London
freshplac.es	http://www.freshplac	We serve you an email digest of the freshest openings near you.	Max S / Tal S	New York City Big Apps 3.0	Javascript	free	Wep App	Eating in NYC	NYC open data	0 tweets + 0 fb likes	NYC citize
Freshly	http://www.hellofresh	Freshly is a mobile app that gives you a new way to look at everything you eat- freshness first. It can tell you when your favorite sushi restaurant has fish delivered, when your local coffee shop roasts beans, or where to find a farmers market to buy fresh produce. It combines posts from local businesses, data sets from NYC Open Data, and the google maps api to let users accurately know what's freshest. You can sort through posts from local NYC businesses based on what's fresh now, most loved by the community, or who you're following. You can also filter by categories, and find out where you're favorite food trucks are going to be today!	various	New York City Big Apps 3.0	iOS Objective C	App Store	Mobile	Eating in NYC	NYC open data , business data	0 tweets + 0 fb likes	NYC citize
My NYC BBQ	http://www.digitalspr	Find out where you can have a BBQ in New York City's Parks. Quick and easy way to view from a list, select city wide or by county, then get a map of location where you can enjoy your BBQ. All you need then are the ribs and sauce! Enjoy Easy to use instructions included but likely not needed.		New York City Big Apps 3.0	N/A Perhaps Android	Free / Digital Spring Apps	Mobile	Eating in NYC	NYC open data , business data	0 tweets + 0 fb likes	NYC citize
NYC Restaurant Explorer	http://resto.davidm.n	Explore an interactive map of NYC to see the diversity of dining choices across the five boroughs.		New York City Big Apps 3.0	Javascript	free	Wep App	Eating in NYC	NYC open data	0 tweets + 0 fb likes	NYC citize
New York Green	http://foodisjoy.org/	NY Green provides information on Green Markets in two forms: a map for currently open markets (and CSAs -Community Supported Agriculture) and a complete list of markets per borough. It displays the complete inventory of all produce grown in the New York area and a time-aware list of the products in season. Products come with a simple seasonal recipe to inspire healthy home-cooking.		New York City Big Apps 3.0	iOS Objective C	App Store	Mobile	Eating in NYC	NYC open data	0 tweets + 0 fb likes	NYC citize

Sample of the List with Open Data Apps metadata yielded from our survey

Open Data Catalogue	Open App Model elements	Comments on Category	Comments on Theme	Comments on technology	General comments	Data Portal
Ottawa Open Data Apps Directory	App Name, Type (mobile, web, desktop), City, Website URL, Author/Published by, Desription, Category (arts, city services, environment, family, finance, government, health & services, housing, infrastructure, maps, parks & recreation, statistics, transportation, other)	46 Apps where 10 pure mobile, 3 pure desktop, 21 pure Web Api and the other combination mainly Web Api / mobile	1 Art , 4 City Services, 3 Environment 1 Family , no finance at all, 3 Goverment, 2 Health and Safety, 2 Infrastracture, 5 maps, 11 parks and recreations, 10 Transportation, 2 miscelleneous, 1 browser for open data and 1 open data catalogue	Most of them used Javascript and google maps or google apis - 1 (the browser) took from the open data katalogue and inserted them to mysql tables so the public can use them		http://www.ottawa.ca/online_services/opendata/info/index_en.html
Data.gov	Description, Activity (Community Rating,Your Rating,Raters, Visits, Downloads, Comments, Contributors), Meta(Category, Permissions, Tags), Links (Perma Link, Short URL), Licensing and Attribution (Data Provided By, Source Link), Dataset Summary (Agency, SubAgency, Date Released, Date updated, Time Period,Frequency,	256 are rss feeds, 858 data extraction tools, 87 widget and gadget	Agriculture 28 Arts, Recreation, and Travel 2 Banking, Finance, and Insurance 7 Births, Deaths, Marriages, and Divorces 6 Business 0 Business Enterprise 54 Census 0 Construction and Housing 15 Contributors 0 Diplomacy 0	Plain data extraction, rss feeds, widget , gadget	1229 goverment apps	
data.gov.au	App Name, Desription, Category (environment, family, finance, government, health , history, housing, infrastructure, maps, parks & recreation, statistics		Many from history,,, and just 1 for transportation. Where are all ther rest ?			http://data.gov.au/data/
New York City Big Apps 3.0	App Name, URL Desription, Category (Developer Tool, Eating in NY, Education, Environment & Sustainability, Exploring NYC, games, Getting Around, Government & Civics, Health & Safety,	26 Web, 57 Mobile (14 Android, 32 IOS, 6 Windows Mobile	4 apps in developing tools, 10 Eating, 4 Education, 8 Environment & Sustainability, 23 Exploring NYC, 2 games, 19 Getting around, 8 Government & Civics, 8 Government & Civics, 10 Living in NYC, 8		Big variety of application	https://nycopendata.socrata.com/
New York City open data goverment apps		944 apps from all the datasets / 186 are maps, 6 are charts, 124 filtered views, 34 files and documents 134 external datasets, 464 datasets	Business and Economic 53 , Community Service 22, Construction and Housing 5, Cultural Affairs 25, Education 112, Environmental Sustainability 51, Events 10, Facilities and Structures 81, Finances 45, Government 47, Health 36, Library 14, Media 14, other 88, Property 42, Public Safety 112, Social Services 39, Statistics 17, Transportation		944 nyc apps	https://nycopendata.socrata.com/
Guardian (open platform)	App Name, URL, Author, Post Date, Category (Politics, Environment, Sport, Music, Data Visualisations, Search, Mobile, Tools)	various	They have stated the more recent 12 Politics, 12 Environment, 8 sport, 3 music, 21 for Data Visualisations, 21 21 for Search Apps, Mobile Apps, 11 Tools	Some apps are only visualization on flickr. They offer their newspaper as app in windows phone, adroid, iphone, blackberry	Guardian has launched OPen Platform The Open Platform is a suite of services that enables partners to build applications with the Guardian. http://www.guardian.co.uk/open-platform/faq	http://www.guardian.co.uk/open-platform/faq
http://data.gov.sg	App Name, Description, PLatform, Government Data used, URL	Transportation				http://data.gov.sg