




Article

The Use of National Strategic Reference Framework Data in Knowledge Graphs and Data Mining to Identify Red Flags

Charalampos Bratsas ^{1,2,*} , Evangelos Chondrokostas ^{1,2} , Kleanthis Koupidis ^{1,2}  and Ioannis Antoniou ^{1,2}

¹ School of Mathematics, Aristotle University of Thessaloniki, 54124 Thessaloniki, Greece; echondrok@gmail.com (E.C.); koupidis.okfgr@gmail.com (K.K.); iantonio@math.auth.gr (I.A.)

² Open Knowledge Foundation Greece, 54352 Thessaloniki, Greece

* Correspondence: cbratsas@math.auth.gr; Tel.: +30-2310-997897

Abstract: Red Flags in fiscal projects are warning signs that may indicate underlying problems with their implementation. In this paper, we present how National Strategic Reference Framework Open Data can be used to take full advantage of semantic web technologies and data mining techniques to build a knowledge-based system that identifies Red Flags. We collected the data from the Open Data API provided by the Greek Ministry of Economy and Finance. Data modeling consist of two ontologies; the Vocabulary of Fiscal Projects, describing the fiscal projects and the National Strategic Reference Framework Greece Vocabulary, illustrating the Greek National Strategic Reference Framework data. We transformed the data into RDF triples and uploaded them onto an OpenLink Virtuoso Server, so that we could retrieve them via SPARQL queries. Performance indicators were defined to assess the state of the project and Density-Based Spatial Clustering of Applications with Noise, (DBSCAN) was used to identify Red Flags. User's demands is that rejected projects should raise Red Flags, to avoid project failure and assist the auditor to organize the monitoring process efficiently, by avoiding to examine most of the non-problematic projects. We performed a use case scenario in which an auditor has to examine NSRF projects, approximately 12 months before the end of the programming period. The system retrieved the fiscal information, calculated the performance indicators and identified the Red Flags. The last update of the projects status after the end of the programming period was retrieved and extracted the number of rejected projects, to test whether the user requirements are satisfied. Rejected projects consist of 3.8% of the total projects. The results of the use case scenario show that *RedFlags* platform is more likely to identify project failures and not raise Red Flags on not rejected projects. Therefore, the *RedFlags* platform using open data, assists the auditor to organize the monitoring process better.

Keywords: Red Flags; Knowledge Graphs; Density Based Clustering; DBSCAN; NSRF Open Data; warning system



Citation: Bratsas, C.; Chondrokostas, E.; Koupidis, K.; Antoniou, I. The Use of National Strategic Reference Framework Data in Knowledge Graphs and Data Mining to Identify Red Flags. *Data* **2021**, *6*, 2. <https://doi.org/10.3390/data6010002>

Received: 6 December 2020

Accepted: 29 December 2020

Published: 4 January 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

A very large amount of the European Union's total budget is spent on regional policy, via the structural funds with the main purpose of reducing the economic disparities between the member states and supporting job creation, business competitiveness, economic growth, sustainable development, and improving the quality of life.

In Greece, the National Strategic Reference Framework (NSRF) establishes the priorities for spending these funds at national level, for a time window of seven years, to raise the competitiveness of the economy, develop human capital and ensure higher employment and income, as well as better social integration [1]. The General Secretariat for Investments and NSRF of Greece, provide online services for access of all interested parties to the NSRF Project Data and to the transparency of the public sector, in accordance with the provisions of Chapter A of the Law 4305/2014 (Government Gazette 237/A) regarding Open disposition and further use of documents, information and data of the public sector [2]. The data

for the NSRF, 2007–2013, are publicly available at <http://2013.anaptyxi.gov.gr>, the official website of the Greek Ministry for Development and Competitiveness, that provides analytical information related to the implementation process of the NSRF projects [3].

According to Fazekas & Tóth [4], EU funds in many cases may increase the risk of corruption and have a negative effect on the development and economic growth of some EU members. During the implementation process of a project unexpected events may occur which could affect project milestones, contracts, payments, or the quality of the product/service being delivered. These events will mark the project as a Red Flag, however the existence of a Red Flag does not necessarily mean that there is corruption in the project [5]. The importance of Red Flags has been indicated in [6–10] and examined with fraud risk indicators related to management fraud [11,12]. However, the results of these studies do not show which fraud risk indicators are the most important [13].

Prior research regarding corruption risk on public procurement include various tools derived from the Subsidystories¹ and Digital Whistleblower² EU project. Subsidystories.eu collects all the data regarding how each country member of the EU allocates its money from the European Structural and Investment Funds (ESIF), making it easier to follow the money. To achieve this goal Subsidystories uses raw data from various portals and official documents from each member state which then visualize by country, in order to trace it easier. The downside is that it is limited on the visualizations and there is no use of data mining to analyze the available data. On the other hand, the Digital Whistleblower project aims to increase fiscal transparency and the impact of good governance policies assessed, through the systematic collection, structuring, analysis, and dissemination of information on public procurement. As part of this project the Monitoring European Tenders (MET)³ risk assessment software was developed. The software supplies the public authorities engaged in procurement activities with an easy-to use tool that can help them identify risky contracts [14].

In addition, another anti-corruption tool that aims to enhance transparency and fight corruption of public procurements is the EU supported project [RedFlags.eu](http://redflags.eu)⁴. This tool is an automatic warning system that uses multiple condition-based algorithms to find Red Flags in the Hungarian procurement documents from Tenders Electronic Daily (TED). Specifically, in their methodology they defined 41 indicators⁵ to monitor the public procurement documents published at the launch and at the end of the procedure. Each indicator has a separate algorithm that raises a Red Flag if certain conditions are met. Many of those algorithms are based on pattern matching techniques or out of range values. In the end, each procurement procedure can potentially have a red flag per indicator and the more flags a procurement has, the riskier it is. Finally, the EU Commission has developed a risk scoring tool called ARACHNE⁶ that performs data mining and data enrichment with the primary objective to support the managing authorities of the member states responsible for EU-funded projects by effectively and efficiently detecting the riskiest projects, contracts, contractors and beneficiaries. ARACHNE just like MET are meant to be used only by public authorities.

In spite of the considerable public and policy interest in corruption and risks in EU Funds spending, citizens, journalists, even public authorities need an open monitoring tool that will identify Red Flags in order to retain transparency policies, take precautionary measures and prevent these warnings from escalating into project failure [15–18].

In terms of data mining algorithms, the DBSCAN algorithm has a variety of data mining uses as it has the ability to handle and identify noise, discover clusters of arbitrary

¹ <http://subsidystories.eu/>

² <http://digiwhist.eu>

³ <https://monitoringeutenders.eu>

⁴ <http://redflags.eu/>

⁵ <http://docs.redflags.eu/developer/engine/gears/indicators/>

⁶ <https://ec.europa.eu/social/main.jsp?catId=325&intPageId=3587&langId=en>

shapes, and automatically discover the number of clusters [19]. DBSCAN is a robust clustering algorithm which has been compared with other data mining algorithms and on a variety of datasets. Recent studies showed that it can be used as part of a system which identified clusters to solve single target and multi-target regression tasks on several datasets [20] and can be used to generate the fault clustering templates for reducing the influence of noise on diagnostic accuracy of rolling bearing datasets. [21]. Additionally, it has been tested on high-dimensional datasets in which clusters are formed by both distance and density structures, where many clustering algorithms fail to identify these clusters correctly [22].

There are various approaches which combine data mining methods and knowledge discovery with Semantic Web data, which support different data mining tasks and improve the Semantic Web [23]. The purpose of this paper is to propose a framework and implement a Knowledge Based system to monitor NSRF projects, using open data and semantic web technologies with linked data principles, to be able to link with other datasets and SPARQL endpoint to retrieve data, performance indicators to monitor the implementation, data mining techniques to identify Red Flags and techniques to visualize the results. This knowledge based system was developed as a web application; *RedFlags*⁷. The rest of the paper is structured as follows: Section 2 provides the complete design of the knowledge based system from the data extraction to the data mining techniques. Section 3 reports the results, Section 4 includes the user requirements and a use case scenario and Section 5 concludes this paper with some directions for future research.

2. Materials and Methods

2.1. Overview

In this section, we describe the knowledge discovery process; the NSRF data used in *RedFlags* application, the vocabularies to semantically represent them, as well as the process for retrieving the needed data using SPARQL queries, defining performance indicators and using data mining techniques to identify Red Flags (Figure 1).

2.2. Data

The official website of the Greek Ministry for Development and Competitiveness publishes data related to the implementation process and the economic activity of the NSRF projects for the programming period at <http://2013.anaptyxi.gov.gr/>. In order to strengthen the transparency of the public sector the database is being updated daily and can be accessed through the Open Data API [24]. These data provide information about two main categories of actions, projects and support-grants.

- Projects: “A group of activities aiming at the realisation of a functionally complete and distinct result. Some projects may consist of other subprojects.” [3].
- Support-Grants: “An advantage in any form whatsoever conferred on a selective basis to organisations involved in economic activity private or public (‘undertakings’) by national public authorities with the potential to distort competition and affect trade between member states of the European Union. The advantage can take different forms of assistance including the direct transfer of resources, such as grants and soft loans, and also indirect assistance, for example, relief from charges that an undertaking normally has to bear, such as a tax exemption or the provision of services, loans, at a favourable rate.” [3].

⁷ <http://redflags.okfn.gr/en/>

ANAPTYXI.gov.gr

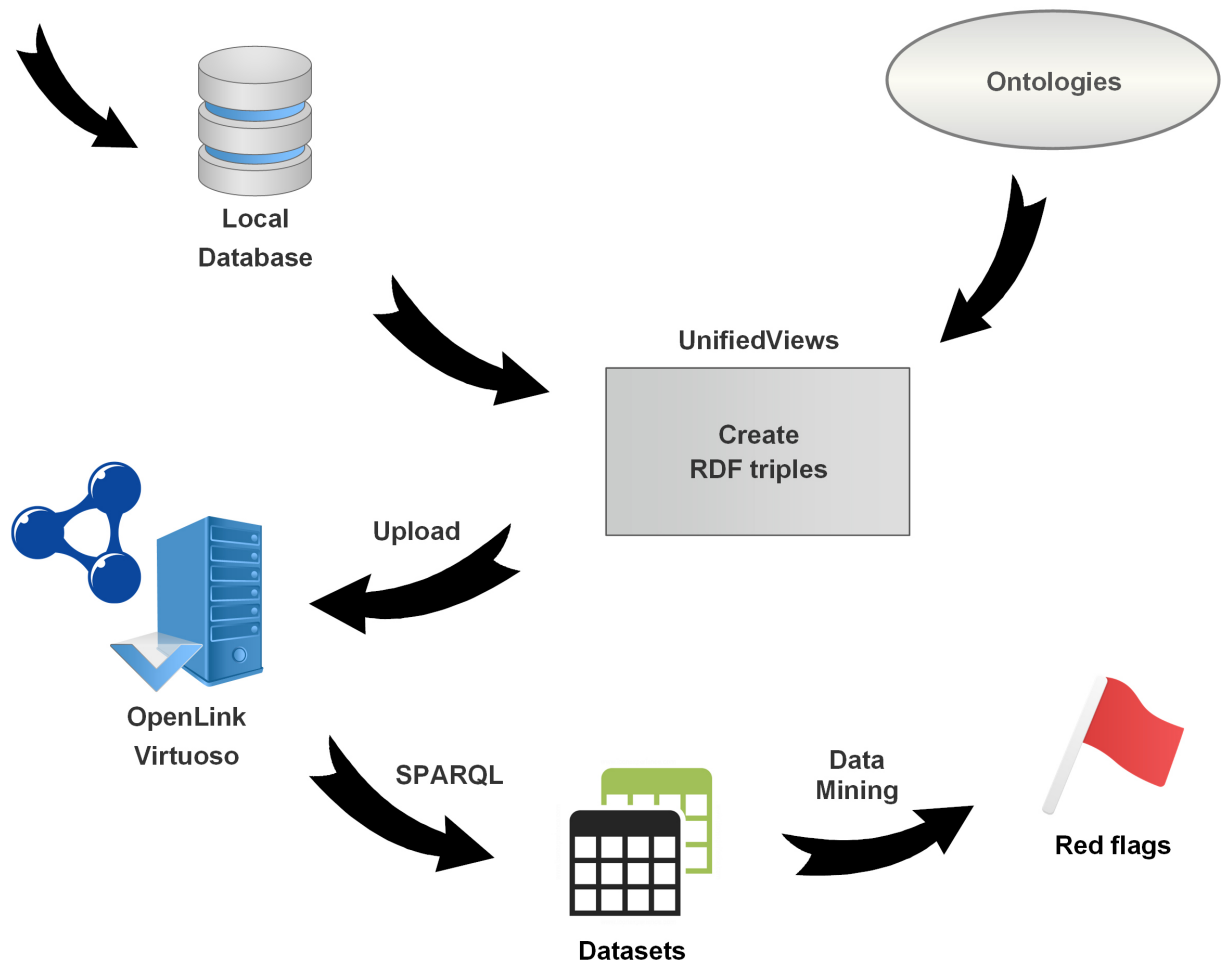


Figure 1. System Architecture.

There is also a category with 181 Priority Projects ...“the selection of which was made by the Greek authorities in cooperation with the qualified European Commission Services, based on criteria related to the maturity, size and importance of their social and economic impact. The Priority Projects consist of other Projects or Support-Grants” [3].

These data include information about the following: public expenditure budget, contracts signed, payment amounts, the start and end date, status, location, description of projects, number and the title of their subprojects, the thematic priority and the operational programme in which they belong, beneficiaries or other involved organisations and various related documents (pictures, pdfs and docs). Also, some projects may involve expropriations. An expropriation is defined as ...“obligatory, according to the law and based on a defined compensation, acquisition of one’s property by the state, for reasons of public necessity or utility” [3]. The expropriation data consist of information about the area, the compensation money and the decisions based on which they are implemented.

2.3. Semantic Data Modeling

Existing vocabularies that could be used to describe fiscal projects and their implementation process are FRAP⁸, an ontology for describing the administrative information of research projects, FP6 and FP7⁹, that were used to model information for European Commission's Framework Programme research projects. These ontologies were very specific about modeling information regarding the research projects and could not be used in our case, which was to describe the properties of a financial project and its implementation process for the Greek NSRF that consists not only of research projects as well infrastructure projects, projects regarding energy, the environment, culture and tourism.

The absence of an ontology that describe financial projects, led us to develop the Vocabulary of Fiscal Projects (VFP) [25] and National Strategic Reference Framework Greece Vocabulary (NSRF-GR) [26] ontologies that could be used as a basis for the semantic representation of the fiscal projects and the Greek NSRF data respectively.

VFP is identified by the namespace URI <http://purl.org/vocab/vfp#>, the preferred prefix is `vfp` and is also available through the GitHub repository¹⁰. The design is based on the research of other EU countries' web portals that provide similar information about projects. Table 1 shows the four main classes we defined to optimise the coverage of terminology in the context of fiscal project data.

The main class of ontology is `vfp:Project`. A project is always associated with some organisations (`vfp:Organization`), a location (`vfp:Place`) and some documents (`vfp:Document`). A more detailed cross reference of the ontology classes and properties is available on its webpage¹¹. Figure 2 depicts the classes and their relations.

Table 1. Classes of the VFP ontology.

Class	Label	Subclass of
<code>vfp:Project</code>	Financial or fiscal Project. It may refer to a construction project or a grant.	<code>foaf:Project</code>
<code>vfp:Organization</code>	Organization related to the project. It includes beneficiaries, contractors/implementers or any other bodies involved in the project.	<code>vcards:Organization</code>
<code>vfp:Document</code>	Documents, images or URLs associated with the project.	<code>foaf:Document</code>
<code>vfp:Place</code>	Place associated with the project.	<code>dbo:Place</code>

NSRF-GR Vocabulary extends VFP with new classes and properties to describe NSRF data in as much detail as possible. It is identified by the namespace URI <http://purl.org/vocab/nsrf-gr#>, the preferred prefix is `nsrf-gr` and is also available through the GitHub repository¹². The classes and its relations are shown in Figure 3. For each project category we created another class, subclass of `vfp:Project`. More details about the classes and the properties can be found at the cross reference section of the ontology's web page¹³.

⁸ <http://www.sparontologies.net/ontologies/frapo>

⁹ <http://mayor2.dia.fi.upm.es/oeg-upm/index.php/en/ontologies/81-research-proj-ontologies/index.html>

¹⁰ <https://raw.githubusercontent.com/okgreece/vfp-ontology/master/vfp.owl>

¹¹ <http://ontologies.okfn.gr/vfp-ontology/index-en.html>

¹² <https://raw.githubusercontent.com/okgreece/nsrf-gr-vocab/master/nsrf-gr.owl>

¹³ <http://ontologies.okfn.gr/nsrf-gr-vocab/index-en.html>

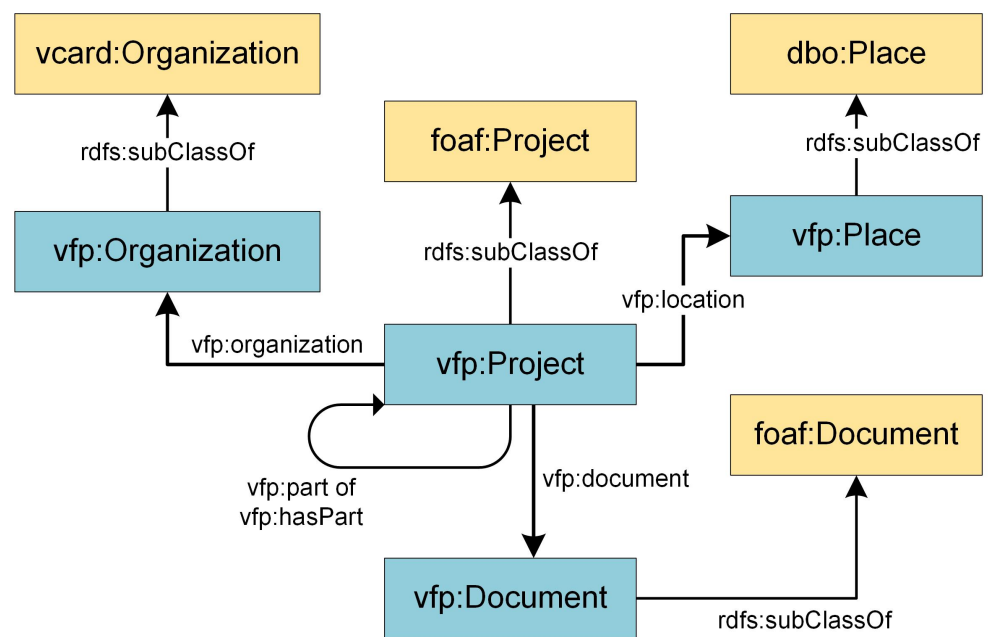


Figure 2. VFP classes and their relations.

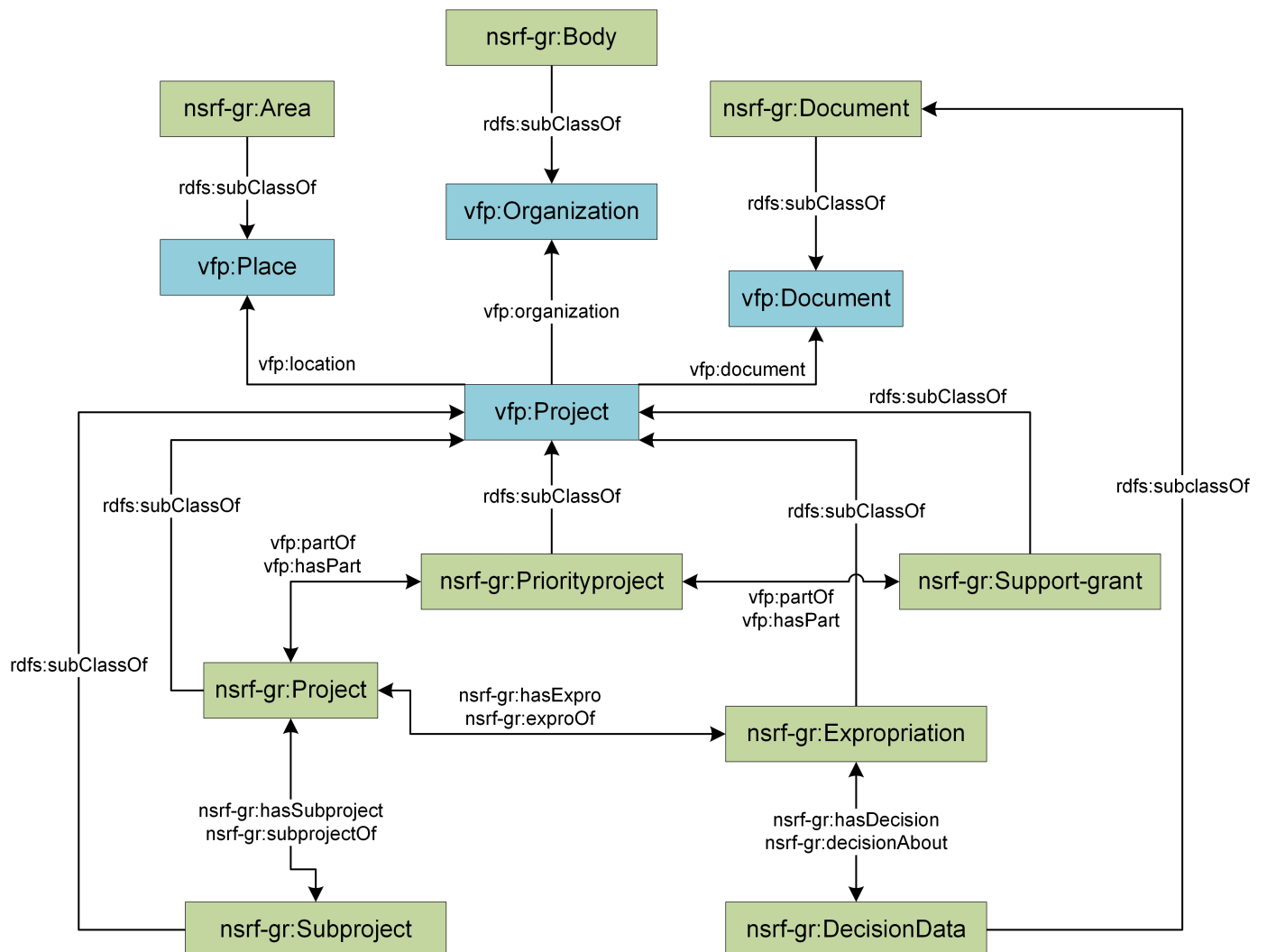


Figure 3. NSRF-GR Vocabulary classes and their relations.

Listing 1. SPARQL query to retrieve basic information about the NSRF projects.

```

PREFIX vfp: <http://purl.org/vocab/vfp#>
PREFIX nsrf-gr: <http://purl.org/vocab/nsrf-gr#>
SELECT ?mis ?title ?description ?body ?cstatus
?location ?operational ?thematic ?url
WHERE {
  ?mis vfp:title ?title .
  ?mis vfp:description ?description.
  ?mis nsrf-gr:body ?body .
  ?mis vfp:currentStatus ?cstatus .
  ?mis vfp:location ?location.
  ?mis nsrf-gr:operational ?operational.
  ?mis nsrf-gr:thematic ?thematic.
  ?mis vfp:url ?url .
}
ORDER BY ?mis
LIMIT 10

```

2.4. NSRF Knowledge Graph and Data Retrieval

We retrieved the data through the Open Data API using Python scripts and stored them in a local database. The transformation of the NSRF data to knowledge graph, is done by using the UnifiedViews¹⁴ ETL tool. The main advantage of this tool is that it can extract data straight from relational databases and then transform it to RDF triples [27–29]. After the transformation process, the RDF files were uploaded to an OpenLink Virtuoso Server¹⁵. Then we used SPARQL queries to retrieve the data from the server and analyze them.

In order to semantically represent the information we extracted from the Open Data Portal about the Greek NSRF projects, we used properties from the VFP ontology to describe the title (vfp:title) of the project, the public expenditure budget (vfp:budget), the total amount of signed contracts (vfp:contracts), the payment amount (vfp:payments), the current status (vfp:currentStatus), the location (vfp:location), a detailed description of the project (vfp:description), its start (vfp:startDate) and end date (vfp:endDate), a status report (vfp:statusReport) and the report date (vfp:statusDate), as well as the url of the project (vfp:url) and the documents related to this project (vfp:document). Also, we used properties from the NSRF-GR vocabulary to represent the project's beneficiary (nsrf-gr:body), the operational programme to which it belongs (nsrf-gr:operational), its thematic priority (nsrf-gr:thematic) and the number of the subprojects it has (nsrf-gr:countSubproject). Finally, all projects have a unique code notated as MIS and were assigned the rdf:type of nsrf-gr:Project.

The object properties vfp:currentStatus, nsrf-gr:operational, nsrf-gr:location, nsrf-gr:body, nsrf-gr:thematic weren't assigned to literal terms, but instead we chose to use code lists. The code lists were semantically represented using SKOS¹⁶, since it's a widespread vocabulary that provides a standard way to organize knowledge using RDF and allows the hierarchical ordering of terms [28].

The query in Listing 1 can be executed on the SPARQL ENDPOINT¹⁷ to retrieve information about the title, description, beneficiary, current status, location, operational programme, thematic priority and the url of the NSRF projects.

¹⁴ <https://unifiedviews.eu/>

¹⁵ <https://virtuoso.openlinksw.com/>

¹⁶ <https://www.w3.org/2004/02/skos/>

¹⁷ <http://data.openbudgets.gr/sparql>

All the IRIs that resulted from the SPARQL query are dereferenceable and point to HTML pages with information about the resources. For the IRI dereferencing we used the RDFBrowser [30], which is an open source Linked Data content negotiator and HTML description generator. Figure 4 shows the HTML representation of the resource project with MIS code 200000.



[Explore Data](#) [Data Toolbox](#) [Documentation](#) [Blog](#) [Login](#)

http://data.openbudgets.gr/resource/dataset/nsrf-gr/mis/200000

A resource of type: nsrf-gr:Project, from Named Graph: http://data.openbudgets.gr/resource/dataset/nsrf-gr/project, within Data Space: http://data.openbudgets.gr/resource/dataset/nsrf-gr/project

Property	Value
nsrf-gr:countSubproject	2 (xsd:int)
vfp:budget	1465906 (xsd:decimal)
vfp:completion	1 (xsd:float)
vfp:contracts	1465906 (xsd:decimal)
vfp:description	ΤΟ ΟΛΟΚΛΗΡΩΜΕΝΟ ΕΡΓΟ ΠΕΡΙΕΛΑΜΒΑΝΕΙ 2 ΥΠΟΕΡΓΑ: - ΥΠΟΕΡΓΟ 1: ΑΝΑΚΑΤΑΣΚΕΥΗ / ΒΕΛΤΙΩΣΗ ΠΑΙΔΙΚΩΝ ΧΑΡΩΝ – ΠΑΙΔΟΤΟΠΩΝ ΤΟΥ ΔΗΜΟΥ - ΥΠΟΕΡΓΟ 2: ΔΗΜΙΟΥΡΓΙΑ ΠΡΑΣΙΝΗΣ ΔΙΑΔΡΟΜΗΣ ΑΠΟ ΤΟ ΠΑΡΚΟ ΠΕΡΙΒΑΛΛΟΝΤΙΚΗΣ ΕΥΑΙΣΘΗΤΟΠΟΙΗΣΗΣ “ΑΝΤΩΝΗΣ ΤΡΙΤΣΗΣ” ΠΡΟΣ ΤΟ ΕΜΠΟΡΙΚΟ - ΔΙΟΙΚΗΤΙΚΟ ΚΕΝΤΡΟ ΤΟΥ ΔΗΜΟΥ ΜΕ ΑΝΑΠΛΑΣΗ ΑΣΤΙΚΟΥ ΧΩΡΟΥ - ΚΑΤΑΣΚΕΥΗ ΠΟΔΗΛΑΤΟΔΡΟΜΟΥ ΚΑΙ ΔΗΜΙΟΥΡΓΙΑ ΥΠΑΙΘΡΙΟΥ ΔΗΜΟΤΙΚΟΥ ΧΩΡΟΥ ΣΤΑΘΜΕΥΣΗΣ ΑΥΤΟΚΙΝΗΤΩΝ ΟΙ ΠΑΡΕΜΒΑΣΕΙΣ ΤΩΝ 2 ΑΥΤΩΝ ΥΠΟΕΡΓΩΝ ΑΦΟΡΟΥΝ: Α) ΣΤΗ ΒΕΛΤΙΩΣΗ ΤΗΣ ΛΕΙΤΟΥΡΓΙΚΟΤΗΤΑΣ ΤΩΝ ΚΟΙΝΟΧΡΗΣΤΩΝ ΧΩΡΩΝ ΚΑΙ ΤΗΝ ΑΝΑΒΑΘΜΙΣΗ ΤΟΥ ΑΣΤΙΚΟΥ ΠΕΡΙΒΑΛΛΟΝΤΟΣ, Β) ΣΤΗ ΒΕΛΤΙΩΣΗ ΤΗΣ ΠΡΟΣΒΑΣΙΜΟΤΗΤΑΣ ΚΑΙ ΤΗΣ ΚΙΝΗΤΙΚΟΤΗΤΑΣ ΣΤΗΝ ΠΟΛΗ ΚΑΙ Γ) ΣΤΟΝ ΕΚΣΥΓΧΡΟΝΙΣΜΟ ΤΩΝ ΠΑΙΔΟΤΟΠΩΝ.
vfp:endDate	2010-12-31 (xsd:date)
vfp:payments	1465906 (xsd:decimal)
vfp:startDate	2009-09-16 (xsd:date)

Figure 4. Example of a project's IRI viewed using the RDFBrowser.

The SPARQL query in Listing 2 can be used to retrieve information about the budget, the contracts, the payments, the start and the end date of the NSRF projects. The results are also shown in Table 2. Data consumers can use the SPARQL ENDPOINT to get information about the Greek NSRF projects, relevant documents, expropriations and their decisions. The SKOSified code lists are also available.

Listing 2. SPARQL query to retrieve fiscal information about the NSRF projects.

```

PREFIX vfp: <http://purl.org/vocab/vfp#>
PREFIX nsrf-gr: <http://purl.org/vocab/nsrf-gr#>
SELECT ?mis ?budget ?contracts ?payments ?startdate ?enddate
WHERE{
?s vfp:budget ?budget ;
  vfp:contracts ?contracts ;
  vfp:payments ?payments ;
  vfp:startDate ?startdate ;
  vfp:endDate ?enddate
  BIND(replace(str(?s),
"http://data.openbudgets.gr/resource/dataset/nsrf-gr/mis/",
"") AS ?mis)
}
ORDER BY ?mis
LIMIT 10

```

Table 2. Raw data retrieved (Budget, Contracts and Payments amounts in €) from the SPARQL query of Listing 2.

MIS	Budget	Contracts	Payments	Start Date	End Date
200000	1,465,906	1,465,906	1,465,906	2009-09-16	2010-12-31
200010	25,346,422	25,346,422	25,346,422	2009-10-12	2016-12-31
200054	6,347,801	6,259,160	5,661,888	2009-03-23	2015-12-31
200056	19,495,000	18,934,124	18,934,124	2009-01-01	2015-11-30
200059	7,011,500	6,817,449	6,801,507	2009-08-14	2015-12-31
200065	9,152,263	9,152,263	9,152,263	2010-12-22	2015-11-30
200101	4,543,729	3,165,602	754,299	2012-07-19	2015-12-31
200111	421,780	421,780	421,780	2010-04-29	2011-06-29
200112	173,720	173,720	173,720	2010-04-01	2011-06-30
200115	55,000	55,000	55,000	2010-09-01	2013-12-31

2.5. Performance Indicators

The process of monitoring and evaluating systems is based on indicators that assess the state of a project [6,7,9,10,13,31]. We use three indicators using the contract, budget and payment amounts from the retrieved data. These indicators track the way in which NSRF projects evolve towards completion and consist of the input features in the clustering algorithm.

The completion index is defined by the Greek Ministry of Economy and Finance as the ratio of payments registered at the moment of data retrieval to the updated budget amount at the moment of data retrieval [3]. We define two other indices, namely, payment completion and contract completion as follows:

Payment completion is defined as the ratio of the payments registered at the moment of data retrieval to the updated contracted amount. The payments completion index shows the status of the payments over the contracts at the time we retrieved the data, while the completion index shows the status of the payments over the budget of the whole project.

Contract completion is the updated contracted amounts to the updated budget at the moment of data retrieval.

The indices range should lie between 0 and 1. A value over 1 means that there is a significant change in a project that was unable to be covered by its fiscal plan and explains why an indicator exceeds the upper limit.

Indicators for each project can be calculated and retrieved using the SPARQL query of Listing 3.

Listing 3. SPARQL query to retrieve indicators for the NSRF projects.

```

PREFIX vfp: <http://purl.org/vocab/vfp#>
PREFIX nsrf-gr: <http://purl.org/vocab/nsrf-gr#>

SELECT ?s ?title
(?payments/?budget AS ?completion)
(?payments/?contracts AS ?payment_completion)
(?contracts/?budget AS ?contract_completion)
WHERE{
?s vfp:title ?title .
?s vfp:budget ?budget .
?s vfp:contracts ?contracts .
?s vfp:payments ?payments .
FILTER (?budget != 0 && ?contracts!=0)
}
ORDER BY ?s

```

2.6. Density Based Clustering

The information if a project is a Red Flag is not available in the official data portal of the Ministry. The available data, described in Section 2.2, concern public expenditure budgets, contracts signed, payment amounts, the start and end date, status, location, description of projects, number and the title of their subprojects, the thematic priority and the operational programme in which they belong, beneficiaries or other involved organisations and various related documents (pictures, pdfs and docs). Supervised approaches are used when we have prior knowledge of what the output values for our samples should be. Therefore, unsupervised learning is appropriate to act on data without categorization [29,32]. Partitioning and hierarchical clustering algorithms are more effective on compact and well separated clusters, however in the presence of noise and outliers in the data, these methods are not very effective [33–35]. We selected Density Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm, to detect areas with high density (clusters of any shape) in the defined feature space (Figure 5) in order to eventually reveal the projects that could be considered as Red Flags.

Having defined a 3-dimensional feature space described in Section 2.5, each project is represented by one point. Let ε the radius of a neighborhood with respect to some point P and $MinPts$ is the minimum number of neighbours within this radius. The notion of density in the feature space is based on the following definitions [19,33]:

- A point P_1 is a core point if at least $MinPts$ points are within distance ε . Those points are said to be directly reachable from P_1 .
- A point P_2 is density reachable to a point P_1 with regard to ε and $MinPts$, if there is a path of core points where each point of the path is directly reachable from the previous one.
- A point P_2 is density connected to a point P_1 with regard to ε and $MinPts$, if there is a point P_3 such that P_1 and P_2 are density reachable from P_3 with respect to ε and $MinPts$.
- A group of density connected points form a density based cluster and points that are not reachable from any other point are outliers.

Based on these density conditions, there are three different kinds of points: core points, density reachable points and outliers, as shown in Figure 5.

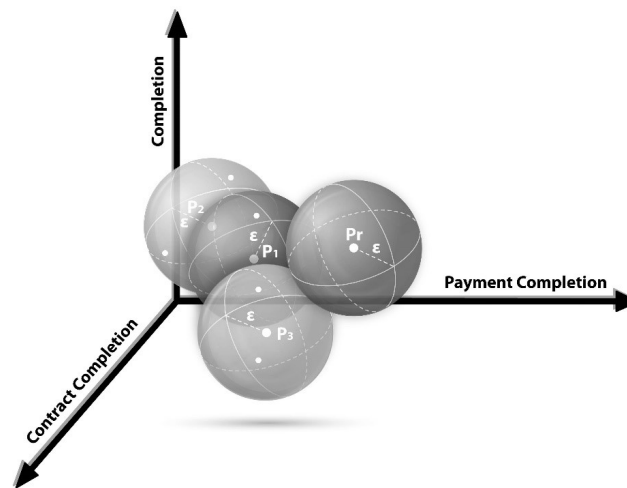


Figure 5. Example of kinds of points DBSCAN uses in 3D feature space. $MinPts = 4$. Points P_1, P_2 are core points, because the area surrounding these points in an ϵ radius contain at least 4 points (including the project itself). Because they are all reachable from one another, they form a single cluster. Point P_3 is not core point, but is reachable from P_1 and thus belongs to the cluster as well. Point P_r is a noise point that is neither a core point nor directly-reachable.

DBSCAN computes the Euclidean distance from an arbitrary selected point (starting point) and the other points and finds the neighbours within their ϵ -distance of the starting one. If the number of neighbours is equal to or greater than the $MinPts$, they form a cluster. These points are considered as “visited”. This process is repeated with the rest core points until the cluster is fully expanded and then, these iterations are also repeated with the unvisited points to form other clusters. If the number of neighbours is less than $MinPts$, the point is marked as a Red Flag.

The rule of thumb, to specify $MinPts$ is to use at least the number of dimensions of the data set plus one. In this case $MinPts$ was set to $k = dim(data) + 1 = 4$ [32,36]. The optimal ϵ radius was specified using a 4-dimensional tree which computes the 4-nearest neighbours’ distances of every point. Figure 6 shows the points sorted by distance in ascending order and the optimal ϵ parameter is selected to be the knee of the curve, the value where a sharp change occurs and is around 0.015 [19].

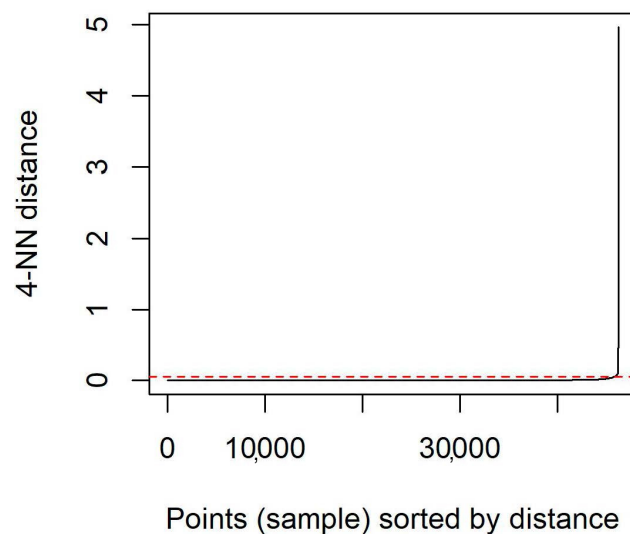


Figure 6. 4-nearest neighbor distance plot.

2.7. Red Flags

Red Flags are defined as clusters of projects with extreme behaviour compared to other clusters of projects. Red Flags are warning signs that do not indicate guilt or innocence [37]. Clusters with a number of projects less than, or equal to 5% of the total number of projects are characterized as extreme clusters. The 5% threshold was selected by trial and error by testing different thresholds.

RedFlags application was built in R (version 3.3.2) [38], with Rstudio (version 1.0.136) [39], using the packages R Shiny (version 1.0) [40], SPARQL (version 1.16) [41], dbSCAN (version 1.0.0) [42], plotly (version 4.5.6) [43], ggplot2 (version 2.2.1) [44], rbokeh (version 0.5.0) [45], DT (version 0.2) [46], shinythemes (version 1.1.1) [47], shinyjs (version 0.9) [48] and shiny-dashboard (version 0.5.3) [49].

3. Results

NSRF data for the programming period 2007–2013 consists of 11,558 projects that were contracted and executed. The proposed performance indicators as retrieved from the SPARQL query (Listing 3) are shown in Table 3.

Table 3. Projects Performance Indicators.

MIS	Completion	Payment Completion	Contract Completion
491704	0.81	0.84	0.97
524889	0.81	1.00	0.81
524944	0.81	1.00	0.81
525053	0.81	0.81	1.00
525097	0.81	0.81	1.00
216685	0.80	1.00	0.80
216686	0.80	1.00	0.80
217143	0.80	1.00	0.80
217183	0.80	0.80	1.00
270967	0.80	1.00	0.80

Table 4 shows the basic descriptive statistics concerning the performance indicators of projects. The large standard deviation indicates the existence of extreme values.

Table 4. Summary Statistics of Performance Indicators.

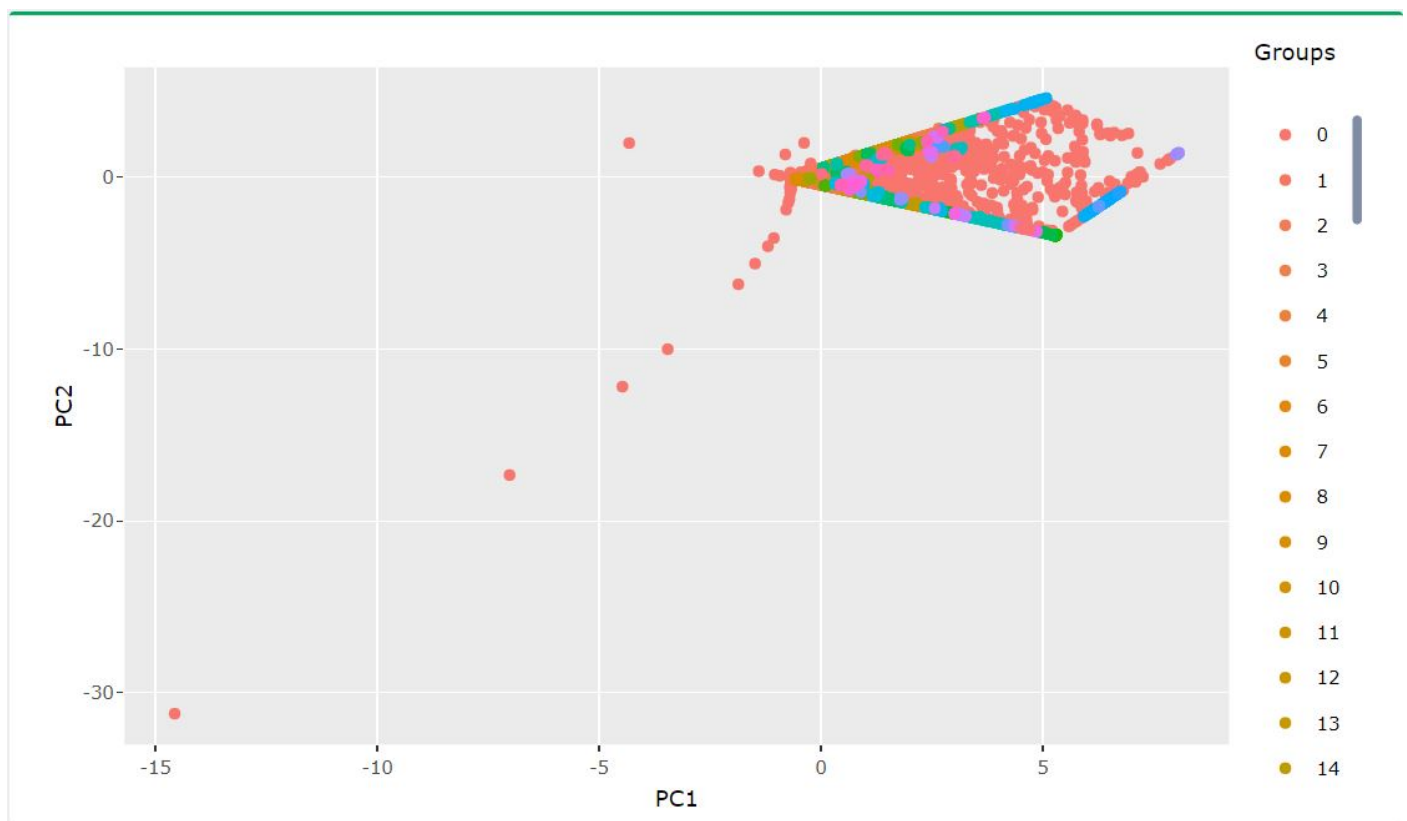
Performance Indicators	Mean	SD	Min	Max
Completion	0.89	0.23	0	1.61
Payment Completion	0.94	0.19	0	1.61
Contract Completion	0.94	0.17	0	6.91

Figure 7 shows how the projects are distributed over the principal components of the feature space. The feature space consisted of completion, payments completion and contract completion. DBSCAN detected areas with high density in the defined feature space and revealed 92 groups of projects. Table 5 shows the ten most populated clusters. Cluster 1 (Figure 7) consists of 8150 projects, a number that exceeds the threshold of 0.05 ($\frac{8150}{11558} = 0.7051 > 0.05$). These projects do not indicate extreme behaviour and have been successfully executed.

The other clusters are Red Flag clusters as they have less than 5% of the total number of projects. The second most populated cluster is cluster 4 and it includes $4.41\% < 5\%$ of the total projects ($\frac{510}{11558} = 0.0441$), the third, cluster 0 consists $4.38\% < 5\%$ of the total projects ($\frac{506}{11558} = 0.0438$) and so forth. In total, 3408 projects were identified as Red Flags consisting of 29.49% ($\frac{3408}{11558} = 0.2949$) of the total projects.

Table 5. Top 10 most populated clusters.

Cluster	Members
1	8150
4	510
0	506
2	286
10	163
3	153
9	139
29	132
6	113
17	107

**Figure 7.** Performance Indicators of projects represented by two principal components.

4. User Requirements and Use Case Scenario

Red Flags are an indication to monitor funded projects during their implementation in order to prevent and guide competent authorities to improve or correct weaknesses or prevent failures in operations, accounts and systems. Therefore, *RedFlags* platform's user requirements are:

1. Rejected projects should raise Red Flags, in order to avert project failure if possible.
2. Assist competent authorities to organize the monitoring process efficiently, without loss or misspend of time, by avoiding to examine most of the non-problematic projects.

According to the user requirements, we performed a use case scenario. In this scenario the competent monitoring authority has to examine NSRF projects, approximately 12 months before the end of the programming period. *RedFlags* platform assists the authority to organize the monitoring process and examine first the projects that raised a Red Flag. Marking a project as Red Flag, means that this project has probably significant problems, such as higher payments than the available budget (completion index), or higher payments than the available contract amounts (contract completion index). These projects have high priority to be examined to avoid rejection. Since the available ground truth is the rejection at the end of the programming period when the data retrieved, we will evaluate the performance of *RedFlags* platform on 438 rejected projects over 11558 NSRF projects. Under these circumstances, the use case scenario will show the performance of the *RedFlags* platform on imbalanced data, since the proportion of rejected projects consist of 3.8% ($\frac{438}{11558} = 0.038$) of the dataset (low prevalence).

The system retrieved the fiscal information, calculated the indicators of the NSRF projects as described in Section 2.5 and identified the Red Flags. To test whether the user requirements are satisfied, we checked the last update of the projects after the end of the NSRF programming period and extracted the number of rejected projects. The following tables show the results of this use case scenario.

The contingency table (Table 6) of rejected projects and projects classified as Red Flags shows that 312 projects raised Red Flag and were rejected (True Positives-TP), 126 projects were rejected and didn't raise Red Flag (False Negatives-FN), 8024 didn't raise Red Flag and were not rejected (True Negatives-TN) and 3096 classified as Red Flags but were not rejected (False Positive-FP). Out of the 11558 projects, 3408 projects were marked as Red Flags.

Table 6. Contingency table of rejected projects and DBSCAN Red Flags (TP = True Positive, FN = False Negative, FP = False Positive and TN = True Negative).

	Rejected	Not Rejected	Total
Red Flag	TP = 312	FP = 3096	3408
No Red Flag	FN = 126	TN = 8024	8150
Total	438	11,120	11,558

According to Table 7, prevalence is equal to 3.8% ($P_r = \frac{TP+FN}{TP+TN+FP+FN} = 0.038$) and is defined as the proportion of rejected projects to the total number of NSRF projects. Low prevalence is expected for a successful NSRF programming period, as a higher percentage of this metric means that the NSRF program encountered problems and that more and more projects failed to complete.

Table 7. Joint probabilities (1 = Rejected, 0 = Not Rejected, r = Red Flag, nr = No Red Flag).

	Rejected	Not Rejected	Total
Red Flag	$P_{r,1} = \frac{312}{11558} = 0.027$	$P_{r,0} = \frac{3096}{11558} = 0.268$	$P_r = 0.295$
No Red Flag	$P_{nr,1} = \frac{126}{11558} = 0.011$	$P_{nr,0} = \frac{8024}{11558} = 0.694$	$P_{nr} = 0.705$
Total	$P_1 = 0.038$	$P_0 = 0.962$	1

By these terms, Precision (Positive Predictive Value-PPV) and Negative Predictive Value (NPV) are equal to 9% ($PPV = \frac{TP}{TP+FP} = 0.09$) and 98% ($NPV = \frac{TN}{TN+FN} = 0.98$), respectively. Precision corresponds to the estimated probability that a project randomly selected from the indicated Red Flags is rejected. Negative Predictive Value corresponds to the probability that a project randomly selected from the set of not indicated projects as Red Flags is not rejected. However, both metrics depend on the prevalence, which in this case is

low and they are not intrinsic to the test, as recall and true negative rate are [50]. The overall accuracy (ACC) of the *RedFlags* platform is equal to 72% ($ACC = \frac{TP+TN}{TP+TN+FP+FN} = 0.72$).

Based on Table 7, which presents the joint probabilities for rejected and Red Flags projects, the conditional probabilities of Table 8 were calculated (see also Figure 8). The results show that recall (Sensitivity, or True Positive Rate-TPR), which is the percentage of raising Red Flags at projects that were rejected after 12 months, is 71% ($TPR = P(r|1) = \frac{P_{r,1}}{P_1} = 0.71$). Recall corresponds to the estimated probability that a project randomly selected from the indicated Red Flags projects will be rejected.

Table 8. Conditional probabilities (1 = Rejected, 0 = Not Rejected, r = Red Flag, nr = No Red Flag).

	Rejected	Not Rejected
Red Flag	$P(r 1) = \frac{P_{r,1}}{P_1} = \frac{0.027}{0.038} = 0.71$	$P(r 0) = \frac{P_{r,0}}{P_0} = \frac{0.268}{0.962} = 0.28$
No Red Flag	$P(nr 1) = \frac{P_{nr,1}}{P_1} = \frac{0.011}{0.038} = 0.29$	$P(nr 0) = \frac{P_{nr,0}}{P_0} = \frac{0.694}{0.962} = 0.72$

Moreover, specificity (SPC) is equal to 72% ($SPC = P(nr|0) = \frac{P_{nr,0}}{P_0} = 0.72$) and is related to the *RedFlags* platform's ability to correctly not raising Red Flags at projects that will not be rejected at the end of the programming period.

In other words, the auditor will not examine first the 72% of the projects that will not be rejected, whereas he will first check the 28% of the projects that will raise a Red Flag but won't be rejected (False Positive Rate-False Alarm), which is satisfactory according to the user's demands. Marking projects as Red Flags does not necessarily mean that these projects will be rejected after 12 months, whereas a project that has been rejected should have raised a Red Flag.

Furthermore, we calculated the Positive likelihood ratio (LR+), Negative likelihood ratio (LR-) and the Diagnostic Odds Ratio (DOR). LR+ is defined as the ratio $\frac{P(r|1)}{P(r|0)} = \frac{0.71}{0.28} = 2.54$. The greater the value of the LR+, the more likely a Red Flag indication is a Red Flag warning for a rejected project. In other words, rejected projects are more likely to raise Red Flags than not rejected, since the ratio is greater than 1. On the other hand, the algorithm avoided an $LR+ < 1$ which would imply that not rejected projects are more likely than rejected projects to receive Red Flags.

LR- is defined as the ratio $\frac{P(nr|1)}{P(nr|0)} = \frac{0.29}{0.72} = 0.40$. The meaning of $LR- < 1$ is that a not rejected project is more likely not to raise a Red Flag than a rejected project. A value greater than 1 would imply that rejected projects are more likely not to raise a Red Flag than not rejected projects.

DOR, which is independent of prevalence, measures the effectiveness of the algorithm. DOR is defined as the ratio of $\frac{LR+}{LR-} = \frac{2.54}{0.40} = 6.35$. The value of DOR is greater than one meaning that the algorithm is discriminating correctly.

Therefore, the *RedFlags* platform user requirements are satisfied. In other words, *RedFlags* is more likely to raise Red Flags on rejected projects and is more likely not to raise a Red Flag on not rejected projects and eventually assist the competent authorities to organize the monitoring process efficiently, by avoiding to examine most of the non-problematic projects.

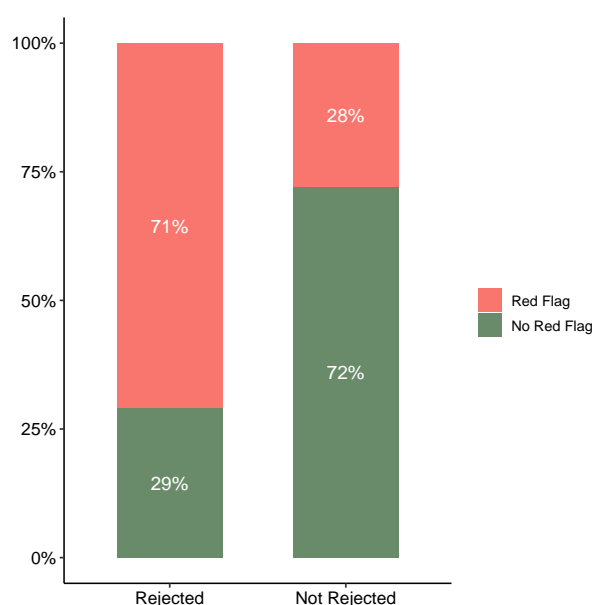


Figure 8. DBSCAN identified 71% of the rejected projects as Red Flags.

5. Conclusions

We presented how open data can be used with semantic web technologies and data mining techniques to identify possible failures as “Red Flags” in National Strategic Reference Framework projects. The identification is implemented by the *RedFlags* application, constructed as an interactive knowledge based system. We used data from the Open Data API provided by the Greek Ministry of Economy and Finance. The semantic description of these data involved the development of two ontologies, VFP and NSRF-GR. The NSRF data were transformed into RDF triples and uploaded to an Openlink Virtuoso Server, while RDFBrowser undertook the process of content negotiation and HTML generation. Performance indicators were defined to track the progress of NSRF projects and provided the inputs to the clustering algorithm. The DBSCAN algorithm was used to identify Red Flags.

The *RedFlags* platform was based on two user requirements. The first requirement is that the rejected projects should raise Red Flags, in order to avoid failure if possible and the second is that there is a need to assist auditors to organize the monitoring process efficiently, without loss or misspend of time, by avoiding to examine most of the non-problematic projects. In the use scenario, an auditor has to examine the NSRF projects in Greece, approximately 12 months before the end of the programming period. The system retrieved the fiscal information, calculated the indicators of the NSRF projects and used the DBSCAN algorithm to identify the Red Flags. *RedFlags* platform marked 29.5% of the projects as Red Flags. The meaning of the indicated Red Flag projects, is that these projects have probably significant problems, due to updates of budget or payment amount, or due to other factors and have high priority to be examined to avoid rejection. However, the available ground truth is the rejection at the end of the programming period when the data retrieved and we evaluated the performance of RedFlags platform on 438 rejected projects over 11558 NSRF projects.

To test whether the user requirements are satisfied, we checked the last update of the projects that were conducted after the end of the NSRF programming period and extracted the number of rejected projects. The number of rejected projects correspond to prevalence which is equal to 3.8% of the total projects. In terms of rejection, low prevalence corresponds to a successful NSRF programming period, as higher values of this metric means that more and more projects failed to complete.

The estimated probability that a project randomly selected from the indicated Red Flags projects will be rejected was 71% (Recall) and the estimated probability to correctly

not raising Red Flags at projects that will not be rejected was 72% (Specificity). Moreover, the positive likelihood ratio showed that rejected projects are more likely than not rejected projects to receive Red Flags, whereas the negative likelihood ratio showed that rejected projects are more likely not to raise a Red Flag than not rejected projects. Finally, the diagnostics odds ratio, which is independent of prevalence, showed that the *RedFlags* platform is discriminating correctly. Therefore, *RedFlags* platform assists the auditor to organize the monitoring process and give high priority at the projects that raised a Red Flag, as rejected projects have higher probability to raise a Red Flag.

Currently the resources in our data have been described by W3C's open standards and have HTTP IRIs so humans can access them and get useful information, but they still don't have links to other datasets. So, our next step will include creating links to IRIs of other published data in order to achieve 5 star Linked Open Data [51,52]. Specifically, we plan to create semantic links between documents that were uploaded to Diavgeia, the official repository where all the decisions of governmental and administrative acts are posted, and the NSRF projects to expand the Greek Linked Open Data (LOD) cloud [53–55]. This will give us access to relevant information about the projects in order to create additional performance indicators and increase the efficiency of the data mining algorithm. In addition, we will further improve the ontology by implementing some upper ontology like BFO¹⁸ and by reusing terms from other ontologies. Moreover, we will look into adding constraints and validating our graphs by using technologies such as SHACL¹⁹ or ShEx²⁰ [56]. Finally, even though the ontologies have their specification drafts, they need to be updated with more detailed documentation and SPARQL examples so consumers, outside of the data portal, can easily compose and execute SPARQL queries using the correct properties.

The findings of this study have been included at the results of the commitment about Linked, Open and Participatory Budgets of the Third Greek Action Plan on Open Government [57]. Public bodies could adapt efficiently to the *RedFlags* Knowledge-Based system as an early warning indicator, in order to make smarter strategies preventing possible failure of projects. Citizens could monitor the progress of a project to find Red Flags, while data journalists could produce data stories about EU funds and relate them with the trends of the Greek economy.

Author Contributions: Conceptualization, C.B., K.K. and I.A.; Data curation, C.B., E.C. and K.K.; Formal analysis, C.B. and K.K.; Investigation, C.B., E.C. and K.K.; Methodology, C.B., E.C. and K.K.; Project administration, C.B.; Software, C.B., E.C. and K.K.; Supervision, C.B. and I.A.; Validation, C.B., K.K. and I.A.; Visualization, C.B., E.C. and K.K.; Writing—original draft, C.B., E.C. and K.K.; Writing—review & editing, C.B., E.C., K.K. and I.A. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been partially supported by the OpenBudgets.eu Horizon 2020 project (Grant Agreement 645833).

Data Availability Statement: The dataset presented in this study is openly available in <http://redflags.okfn.gr/> and can be retrieved from <http://data.openbudgets.gr/sparql>.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ministry of Economy and Finance and General Secretariat for Investments and Development. National Strategic Reference Framework 2007–2013. Available online: http://2007-2013.espa.gr/elibrary/NSRF%20document_english.pdf (accessed on 12 March 2018).

¹⁸ <https://basic-formal-ontology.org>

¹⁹ <https://www.w3.org/TR/shacl/>

²⁰ <https://shex.io/>

2. Hellenic Parliament. Open Disposition and Further Use of Documents, Information and Data of the Public Sector, Amendment of Law 3448/2006 (A'57), Adjustment of the National Legislation to the Provisions of Directive 2013/37/EU of the European Parliament and of the Council, Further Strengthening of Clarity, Regulations of Issues Concerning the Entrance Competition to the National School of Public Administration and Local Government and Other Provisions. 2014. Available online: http://www.hellenicparliament.gr/en/Nomothetiko-Ergo/Anazitisi-Nomothetikou-Ergou?law_id=300b8ca3-3468-4893-9336-9973d3fa247d (accessed on 12 March 2018).
3. Ministry of Economy & Development. What is ANAPTYKSI.gov.gr. Available online: <http://2013.anaptyxi.gov.gr/Default.aspx?tabid=249&language=en-US> (accessed on 12 March 2018).
4. Fazekas, M.; Tóth, I.J. 13 Corruption in EU Funds? Europe-wide evidence on the corruption effect of EU funded public contracting. Available online: http://real.mtak.hu/80734/1/10.4324_9781315401867_14_u.pdf (accessed on 3 January 2021).
5. Kenny, C.; Musatova, M. 'Red Flags Of Corruption' In *World Bank Projects : An Analysis Of Infrastructure Contracts*; The World Bank: Washington, DC, USA, 2010. [CrossRef]
6. Apostolou, B.A.; Hassell, J.M.; Webber, S.A.; Sumners, G.E. The Relative Importance of Management Fraud Risk Factors. *Behav. Res. Account.* **2001**, *13*, 1–24. [CrossRef]
7. Hackenbrack, K. The effect of experience with different sized clients on auditor evaluations of fraudulent financial reporting indicators. *Auditing* **1993**, *12*, 99.
8. Loebbecke, J.; Eining, M.; Willingham, J. Auditorsexperience with material irregularities: Frequency, nature, and detectability. *Auditing: Frequency. J. Pract. Theory* **1989**, *9*, 1–28.
9. Majid, A.; Gul, F.A.; Tsui, J.S.L. An Analysis of Hong Kong Auditors' Perceptions of the Importance of Selected Red Flag Factors in Risk Assessment. *J. Bus. Ethics* **2001**, *32*, 263–274. [CrossRef]
10. Mock, T.J.; Turner, J.L. Auditor Identification of Fraud Risk Factors and their Impact on Audit Programs. *Int. J. Audit.* **2005**, *9*, 59–77. [CrossRef]
11. Coram, P.; Ferguson, C.; Moroney, R. Internal audit, alternative internal audit structures and the level of misappropriation of assets fraud. *Account. Financ.* **2008**, *48*, 543–559. [CrossRef]
12. Liou, F. Fraudulent financial reporting detection and business failure prediction models: A comparison. *Manag. Audit. J.* **2008**, *23*, 650–662. [CrossRef]
13. Gullkvist, B.; Jokipii, A. Perceived importance of red flags across fraud types. *Crit. Perspect. Account.* **2013**, *24*, 44–61. doi:10.1016/j.cpa.2012.01.004. [CrossRef]
14. Calderoni, F.; Milani, R.; Rotondi, M.; Carbone, C.; Savona, E.; Riccardi, M.; Mancuso, M. Public procurement risk assessment software for authorities, 2018. Deliverable 4.4. of the DIGIWHIST project funded under the European Union's Horizon 2020 research and innovation Programme under the G.A. No: 645852. Available online: <https://digiwhist.eu/wp-content/uploads/2018/02/4.4-MET-Risk-Assessment-tool.pdf> (accessed on 24 September 2019).
15. Dimulescu, V.; Pop, R.; Doroftei, I.M. Risks of corruption and the management of EU funds in Romania. *Rom. J. Political Sci.* **2013**, *13*, 101–123.
16. Beblavý, M.; Šiřáková-Beblavá, E. The Changing Faces of Europeanisation: How Did the European Union Influence Corruption in Slovakia Before and After Accession? *Eur.-Asia Stud.* **2014**, *66*, 536–556. [CrossRef]
17. Fazekas, M.; Chvalkovska, J.; Skuhrovec, J.; Tóth, I.J.; King, L.P. Are EU funds a corruption risk? The impact of EU funds on grand corruption in Central and Eastern Europe. *Anticorruption Frontline ANTICORRP Proj.* **2013**, *2*, 68–89. [CrossRef]
18. Hajek, P.; Henriques, R. Mining corporate annual reports for intelligent detection of financial statement fraud – A comparative study of machine learning methods. *Knowl.-Based Syst.* **2017**, *128*, 139–152. doi:10.1016/j.knosys.2017.05.001. [CrossRef]
19. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining KDD'96*, Portland, OR, USA, 2–4 August 1996; AAAI Press: Menlo Park, CA, USA, 1996; pp. 226–231.
20. Corizzo, R.; Pio, G.; Ceci, M.; Malerba, D. DENCAST: distributed density-based clustering for multi-target regression. *J. Big Data* **2019**, *6*. [CrossRef]
21. Li, H.; Wang, W.; Huang, P.; Li, Q. Fault diagnosis of rolling bearing using symmetrized dot pattern and density-based clustering. *Measurement* **2020**, *152*, 107293. doi:10.1016/j.measurement.2019.107293. [CrossRef]
22. Thrun, M.; Ultsch, A. Using Projection-Based Clustering to Find Distance- and Density-Based Clusters in High-Dimensional Data. *J. Classif.* **2020**. [CrossRef]
23. Ristoski, P.; Paulheim, H. Semantic Web in data mining and knowledge discovery: A comprehensive survey. *J. Web Semant.* **2016**, *36*, 1–22. doi:10.1016/j.websem.2016.01.001. [CrossRef]
24. Ministry of Economy & Development. Open Data API ANAPTYXI.gov.gr. Available online: <http://2013.anaptyxi.gov.gr/Default.aspx?tabid=251&language=en-US> (accessed on 12 December 2020).
25. Chondrokostas, E.; Bratsas, C. Vocabulary of Fiscal Projects: VFP v1.1.0. 2019. Available online: <https://doi.org/10.5281/zenodo.3242356> (accessed on 3 January 2021).
26. Chondrokostas, E.; Bratsas, C. National Strategic Reference Framework Vocabulary: NSRF-GR v1.1.0. 2019. Available online: <https://doi.org/10.5281/zenodo.3242355> (accessed on 3 January 2021).

27. Knap, T.; Hanecák, P.; Klímek, J.; Mader, C.; Necaský, M.; Nuffelen, B.V.; Skoda, P. UnifiedViews: An ETL tool for RDF data management. *Semant. Web* **2018**, *9*, 661–676. [CrossRef]
28. Filippidis, P.; Karampatakis, S.; Koupidis, K.; Ioannidis, L.; Bratsas, C. The code lists case: Identifying and linking the key parts of fiscal datasets. In Proceedings of the 2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Thessaloniki, Greece, 20–21 October 2016; pp. 165–170. [CrossRef]
29. Koupidis, K.; Bratsas, C.; Karampatakis, S.; Martzopoulou, A.; Antoniou, I. Fiscal Knowledge discovery in Municipalities of Athens and Thessaloniki via Linked Open Data. In Proceedings of the 2016 11th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Thessaloniki, Greece, 20–21 October 2016; pp. 171–176. [CrossRef]
30. Karampatakis, S. RDFBrowser: An Open Source Linked Data Content Negotiator and HTML Description Generator. Available online: <https://github.com/okgreece/RDFBrowser> (accessed on 12 December 2020).
31. Smith, M.; Haji Omar, N.; Iskandar Zulkarnain Sayd Idris, S. and Baharuddin, I Auditors' perception of fraud risk indicators: Malaysian evidence. *Manag. Audit. J.* **2005**, *20*, 73–85. [CrossRef]
32. Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *KDD* **1996**, *96*, 34.
33. Kassambara, A. *Practical Guide to Cluster Analysis in R: Unsupervised Machine Learning*; STHDA: Montpellier, France, 2017; Volume 1.
34. Xie, J.; Xiong, Z.Y.; Zhang, Y.F.; Feng, Y.; Ma, J. Density core-based clustering algorithm with dynamic scanning radius. *Knowl.-Based Syst.* **2018**, *142*, 58–70. doi:10.1016/j.knosys.2017.11.025. [CrossRef]
35. Zhou, Z.; Si, G.; Zhang, Y.; Zheng, K. Robust clustering by identifying the veins of clusters based on kernel density estimation. *Knowl.-Based Syst.* **2018**, *159*, 309–320. doi:10.1016/j.knosys.2018.06.021. [CrossRef]
36. Sander, J.; Ester, M.; Kriegel, H.P.; Xu, X. Density-Based Clustering in Spatial Databases: The Algorithm GDBSCAN and Its Applications. *Data Min. Knowl. Discov.* **1998**, *2*, 169–194. [CrossRef]
37. Office of the State Comptroller, State of New York. Red Flags for Fraud. Available online: https://www.osc.state.ny.us/localgov/pubs/red_flags_fraud.pdf. (accessed 12 December 2020).
38. R Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2019.
39. RStudio Team. *RStudio: Integrated Development Environment for R*; RStudio, Inc.: Boston, MA, USA, 2018.
40. Chang, W.; Cheng, J.; Allaire, J.; Xie, Y.; McPherson, J. Shiny: Web Application Framework for R. Available online: <https://cran.r-project.org/web/packages/shiny/index.html> (accessed on 12 December 2020).
41. Van Hage, W.R.; Tomi, K.; Graeler, B.; Davis, C.; Hoeksema, J.; Ruttenberg, A.; Bahls, D. SPARQL: SPARQL client. Available online: <https://cran.r-project.org/web/packages/SPARQL/SPARQL.pdf>. (accessed on 12 December 2020).
42. Hahsler, M.; Piekenbrock, M. dbscan: Density Based Clustering of Applications with Noise (DBSCAN) and Related Algorithms. Available online: <https://github.com/mhahsler/dbscan> (accessed on 12 December 2020).
43. Sievert, C. *plotly for R*; R Foundation for Statistical Computing: Vienna, Austria, 2016.
44. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*; Springer: New York, NY, USA, 2016.
45. Hafen, R.; Continuum Analytics, Inc. *rbokeh: R Interface for Bokeh*; R package version 0.5.0; R Foundation for Statistical Computing: Vienna, Austria, 2016.
46. Xie, Y.; Cheng, J.; Tan, X. DT: A Wrapper of the JavaScript Library 'DataTables'; R package version 0.2; R Foundation for Statistical Computing: Vienna, Austria, 2016.
47. Chang, W. *Shinythemes: Themes for Shiny*; R package version 1.1.1; R Foundation for Statistical Computing: Vienna, Austria, 2016.
48. Attali, D. *Shinyjs: Easily Improve the User Experience of Your Shiny Apps in Seconds*; R package version 0.9; R Foundation for Statistical Computing: Vienna, Austria, 2016.
49. Chang, W.; Borges Ribeiro, B. *Shinydashboard: Create Dashboards with 'Shiny'*; R package version 0.5.3; R Foundation for Statistical Computing: Vienna, Austria, 2016.
50. Altman, D.G.; Bland, J.M. Statistics Notes: Diagnostic tests 2: predictive values. *BMJ* **1994**, *309*, 102. [CrossRef] [PubMed]
51. Berners-Lee, T. Design Issues: Linked Data. 2006. Available online: <http://www.w3.org/DesignIssues/LinkedData.html> (accessed on 12 December 2020).
52. Bizer, C.; Heath, T.; Berners-Lee, T. Linked Data—The Story So Far. *Int. J. Semantic Web Inf. Syst.* **2009**, *5*, 1–22. [CrossRef]
53. Bratsas, C.; Alexiou, S.; Kontokostas, D.; Parapontis, I.; Antoniou, I.; Metakides, G. Greek Open Data in the Age of Linked Data: A Demonstration of LOD Internationalization. Available online: <https://doi.org/10.2139/ssrn.2088076> (accessed on 3 January 2021).
54. Kontokostas, D.; Bratsas, C.; Auer, S.; Hellmann, S.; Antoniou, I.; Metakides, G. Internationalization of Linked Data: The case of the Greek DBpedia edition. *J. Web Semant.* **2012**, *15*, 51–61. doi:10.1016/j.websem.2012.01.001. [CrossRef]
55. Kontokostas, D.; Bratsas, C.; Auer, S.; Hellmann, S.; Antoniou, I.; Metakides, G. Towards linked data internationalization-realizing the greek dbpedia. In Proceedings of the ACM WebSci'11, Koblenz, Germany, 14–17 June 2011.

-
56. Gayo, J.E.L.; Prud'hommeaux, E.; Boneva, I.; Kontokostas, D. Validating RDF Data. *Synth. Lect. Semant. Web Theory Technol.* **2017**, *7*, 1–328. [[CrossRef](#)]
 57. Melidis, A.; Deligiannis, A.; Priftis, A. *Greece End-of-Term Report 2016-2018 Open Government Partnership; Independent Reporting Mechanism (IRM)*: Athens, Greece, 2019; pp. 70–71.