INTERNATIONAL
HELLENIC
UNIVERSITY

Extracting Structured Information From Diavgeia Portal

Eleni Tsironidou

SID: 3308180024

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

23th October 2020

THESSALONIKI – GREECE

Extracting Structured Information From Diavgeia Portal

Eleni Tsironidou

SID:3308180024

Supervisor: Professor Berberidis Christos

SCHOOL OF SCIENCE & TECHNOLOGY

A thesis submitted for the degree of

Master of Science (MSc) in Data Science

23th October 2020

THESSALONIKI – GREECE

Abstract

Today’s overload of information, particularly through the World Wide Web, makes it difficult for users to access the right information.Over the last years, there has been an exponential increase in the amount of information stored on the Web. However, most of the information comes in an unstructured format making it difficult to easily and automatically extract knowledge. Therefore, there is a need to extract structured information to increase transparency and create data with high quality that can be further visualized and analyzed to extract insights and knowledge. Recently, machine learning and natural language processing techniques, like named entity recognition and relation extraction,have shown promising results on information extraction.However, current research is mostly focused on the analysis of content in the English language. In this thesis, we worked on extracting structured information from Diavgeia portal, which is a Greek transparency initiative and platform that stores all the administrative acts and decisions made in the public sector. More specifically, we studied the existing solutions for information extraction on Greek documents and performed a case study on Diavgeia portal. Currently, decisions in Diavgeia are uploaded in PDF format with limited meta-data information, being far away from Tim Berners-Lee suggestion on linked open data. Focusing on transparency, we also provided visualizations and statistics on the extracted structured information that could help the users to easily understand more information on Diavgeia. The analysis has shown that using information extraction techniques can significantly improve data quality and subsequently transparency by using more advanced visualizations and search capabilities. However, there is a need for implementation of customized named entity recognition and relation extraction solutions on Greek decisions to further enhance the data quality.

Acknowledgements

At this point I would like to personally thank my supervisor Dr. Christos Berberidis for his valuable help and instructions throughout this work. Moreover, for his guidance instructions during this works as well as his significant support and patience. I would like also to thank Mr. Ioannis Konstantinidis for his co-work with me and his helpful advice for every piece of my master thesis. Finally, I must thank my MSc in Data Science fellow, George Patrikios, for his support and continuous encouragement.

Tsironidou Eleni

04/01/2021

Contents

ABSTRACT 3

1 INTRODUCTION **7**

2 RELATED WORK **10**

2.1 E-GOVERNMENT DATA INITIATIVES 10

2.2 INFORMATION EXTRACTION ON GREEK LEGAK DOCUMENTS USING NLP 11

3 BACKGROUND **14**

3.1 BACKGROUND ON DIAVGEIA 14

3.1.1 Open Data *14*

3.1.2 Diavgeia Portal *14*

3.1.3 Metadata of Greek Legislation *16*

3.2 NAMED ENTITY RECOGNITION 18

3.3 KNOWLEDGE GRAPHS 20

4 MATERIAL AND METHODS **21**

4.1 DATA MODEL 21

4.2 COMMUNICATION WITH DIAVGEIA PORTAL 23

4.3 ANALYSIS METHODS 24

4.4.1 Data Parsing *24*

4.4.2 Data set preprocessing *26*

4.4.3 Entity recognition *28*

4.4.4 Graph Implementation *29*

4.4.4.1Networkx *29*

4.4.4.2Gephi *29*

4.4.5 Network measures *30*

5 EXPERIMENTAL RESULTS **35**

5.1 DATA SET ANALYSIS 35

5.2 NETWORK AND GRAPH FASHIONING 36

5.3 GRAPH ANALYSIS 36

5.3.1Graph visualization *36*

5.3.2 Statistical analysis *38*

6 DISCUSSION... 44

7 CONLUSIONSAND FUTURE WORK 46

BIBLIOGRAPHY... 47

1 Introduction

Over the last years, there has been an increasing interest in open government. Open government is “the governing doctrine which holds that citizens have the right to access the documents and proceedings of the government to allow for effective public oversight”[35]. It could provide considerable benefits for society (like economic growth, stimulation of innovation, promotion of transparency and increased accountability). A special type of public government data that originates from the operation of the three branches of Government (executive, legislative and judicial) is the administrative acts and decisions that come in the unstructured form of PDF documents. Furthermore, due to the rapid increase of documents’ publication, open e-Government data follows big data characteristics with high volume, variety and velocity. These facts constitute an impediment to transparency and search ability over public administrative acts and decisions. According to Tim Berners-Lee[1] [1], there are five levels of data quality:

- 1 star: Data is available on the Web, in whatever format (e.g. PDF).
- 2 stars: Data is available as machine-readable structured data, in whatever format (e.g. Excel file).
- 3 stars: Data is available as machine-readable structured data, in an open, non-proprietary format, (e.g., CSV, XML).
- 4 stars: Data is published using open standards from the W3C (RDF and SPARQL).
- 5 stars: All of the above and links to other Linked Open Data.

Therefore, there is a need to extract structured information from these documents to create high-quality data that the users can further analyze and visualize to gain more insights.

Manually encoding these constantly growing sets of public documents to a structured format is a laborious process. As a result, their manual transformation into Legal Open Data,

[1] ★ OPEN DATA, https://5stardata.info/en/

according to the 5-star deployment scheme proposed by Tim Berners-Lee, seems to be an unrealistic approach (PDF documents are 1 star, meaning the most unstructured form). On the other hand, applying even semi-automatic approaches for the markup of legal texts may greatly reduce the time and effort needed compared to manual text annotation

Recently, innovative advances in the zone of the semantic Web have offered significant improvement towards the Web of Data. The research focuses on linked open data concentrates on how information that is communicated in PDF will be accessible on the Web and linked with other information with the goal of expanding its incentive for everyone. Along these lines, semantic Web technologies give much advancement to making government information open.

A significant sort of government information is the identified with administrative acts and decisions. Legislation applies to each part of individuals’ living and develops constantly constructing a gigantic organization of interlinked authoritative archives. Subsequently, it is significant for an administration to offer administrations that make judicial decisions effectively available to people in general targeting illuminating them, empowering them to protect their privileges, or to utilize legislation as an aspect of their responsibilities.

Apart from the transformation of legal documents to public open data, Information Extraction (IE) task has been taken into serious consideration in order to extract useful data from the documents and also to increase those data quality. Various Natural Language Processing (NLP) tools are developed and used extensively for IE, like spacy, NLTK, CLTK. IE approaches in the legal domain are considerably different from other knowledge areas because of the two main characteristics of legal texts. Legal documents exhibit a wide range of internal structure, and they often have a significant amount of manual editorial value added. Moreover, information extraction from texts written in languages except from the English language is not much effective. Thus, other modern approaches based on combinations of lists, rules, and supervised machine learninghasbeendeveloped.

Towards this bearing, there are as of now numerous European Union (EU) nations that have mechanized the authoritative cycle by creating stages for filing decision records and offering on-line admittance to them. In Greece, up until this point, there has been a restricted level of computerization of the administrative cycle and even the disclosure of administrative acts and decisions identified with a specific theme can be a hard undertaking. The administrative work of the Greek government has been distributed since 1907 as a journal by the National Printing Office .Decisions are distributed consistently in that newspaper and it is conveyed distinctly in a PDF design. An important initiative for making Greek acts more open to people, in general, has been made by Diavgeia[2] , a Greek program presented in 2010, upholding straightforwardness over government and policy implementation by necessitating that administration and policy management need to transfer their decisions on the Web. In any case, no basic design is authorized nor any organizing of their printed content.

In this thesis, our aim is to extract structured information from Diavgeia portal that not only present the textual content of legal documents but also to transform this piece of information to a format that is publicly accessible and transparent and can be further analyzed and visualized. Specifically, we perform a task called Named Entity Recognition aiming on gathering entities like persons, organizations and locations. Moreover, we apply all the information we collect to create a knowledge graph and finally we provide a statistical analysis of the knowledge graphand also visualization results by exploiting the extracted structured information.

The rest of the thesis is organized as follows: Chapter 2 sets the background on information extraction in terms of legislation documents by providing on overview of past research studies regarding the definition of open data andthe approaches of information extraction of Greek legal documents. Chapter 3 presents the background information in terms of Diavgeia portal, the named entity recognition and knowledge graphs theory. Chapter 4 presents the methodology followed in order to connect with Diavgeia API and extract legal documents. Moreover it displays the procedure of how the collected files formed into a data set which is also described. Finally in this chapter we describe the network metrics which are going to be used for the network analysis. Chapter 5 presents the results of the named entity recognition and the graph implementation. In chapter 6 we discuss the findings of our research and detect some limitations. Finally, chapter 7 gives a conclusion of what we have achieved and set some future direction to overcome the limitations.

[2] https://diavgeia.gov.gr/

2. Related Work

In this chapter there will be highlighted past research on the effort of transformation of legislation with unstructured format to public open data following a format being able to be read from citizens. There will provide also the effort of extracting useful information and metadata from legal documents in order to give public the opportunity to not only have access to government decisions but also to build an environment of transparency and integrity between the government and the public sector.

2.1 E-Government Data Initiatives

European citizens and entrepreneurs are increasingly faced with rules and regulations affecting various aspects of their daily activities. These provisions come from foreign, European, national and local authorities. Despite attempts at harmonization and modernization, the amount and size of the theoretically applicable law body is growing. This is also a concern for government, both legislative and executive entities. The process of establishing consistent and coherent legislation is becoming more complex, as it is the procedure of enforcing and introducing valid legislation. ICT has the power to help both the government and the people in dealing with this increasing legal framework. In order to make the law readily available to the public, the creation of information systems archiving legal material and metadata has become a standard practice. A necessary prerequisitefor this is the electronic availability of legal source in a structured and standard format.

For few examples, the Dutch national regulations released by the official site of the Dutch government are provided on the MetaLex document server which is a project to enhance access to legal sources (regulations, court rulings) through a standardized legal XML syntax (CEN MetaLex) and Linked Data. [2]

The MDS describes a generic conversion process from legacy legal XML syntax to CEN MetaLex, RDF and Pajek network files and discloses content through HTTP- based content negotiation, SPARQL endpoint and simple search interface. MDS enhances the transparency of (Modified) and ambiguous (content- based) naming scheme for URIs of sections of legal documents, enabling the monitoring of version information at URI level, as well as reverse engineering of versioned metadata from sources that provide only partial information, such as many web-based legal content services. The MetaLex is the host of all 28k of the Dutch national regulations available since May 2011 and contains some 100M triples.[3]

Another project for the publishing of legislative documents is Finlex Data Bank1 [4] which have been released by the Finnish government for the search and accessing of legislative information. Finlex is based on XML schema and does not comply with modern semantic metadata standards. The Publications Office of the EU has created a central repository of content and metadata, called CELLAR, for the storage of the official publications and bibliographic resources provided by the EU institutions[5], while the United Kingdom publishes laws on its official website.

In Greece, up to this point, there has been a small degree of computerization of the administrative process and indeed the revelation of enactment related to a certain topic can be a difficult and challenging task. The authoritative work of the Greek government has been published since 1907 within the frame of a gazette by the National Printing Office. The law is published regularly in the gazette and it's only available in PDF format. A first attempt to make Greek legislation available to the general public is enacted by Diavgeia, a Greek program launched in 2010 to ensure accountability over government and public by requiring that government and public need to transfer their decisions on internet. Nevertheless, no standard format or any structuring of its textual content is implemented. The platform is operated by the Ministry of Administrative Reform and E-Government. Diavgeia has already been completely enforced by the public authorities. The current rate of uploads within the Diavgeia site is 16,000 decisions per working day, reaching so far the total amount of 26 million posted laws .

2.2 Information extraction on Greek legal documents using NLP

Many of the above activities are focused on semantic wed technology for designing, querying and making policy content readily available to the public. The implementation of the web benchmarks like XML, RDF[3], SPARQL as well as common hones for distributing such

[3] https://en.wikipedia.org/wiki/Resource_Description_Framework

information connected information has been a common practice among these efforts within the plan of vocabularies and ontologies (OWL)[4]for legislative records.

One such vocabulary is Akoma Ntoso (*A*rchitecture for *K*nowledge-*O*riented *M*anagement of *A*frican *N*ormative *T*exts using *O*pen *S*tandards and *O*ntologies)[5] is an international technical standard for representing executive, legislative and judiciary documents in a structured manner using a domain specific, legal XML schema which was sponsored by the United Nations[6][7]. Another is MetaLex, which was initially proposed as an XML vocabulary for the representation of legislative documents' structure and text, but later modified with timekeeping and version control features.[8][9] The MetaLex vocabulary which has been adopted from European Committee from Standardization (CEN) and later developed into an ontology was named CEN Metalex. Even more recently, the European Council presented a framework aimed at consolidating and linking national legislation with European legislation, called ELI (European Legislation Identifier)[10]. As a context, ELI offers a URI format of legal assets on the internet and it moreover gives an OWL philosophy, which is utilized for communicating metadata of legislative documents and activities. ELI, like Akoma Ntoso and MetaLex, isn't a one-size-to-all show but it needs to be expanded to capture the nuances of national frameworks of legislation.

Nevertheless, other practices have already adopted the principles of these vocabularies for the legal domain implementation of appropriate activities. For example, LegalDocML and LegalRuleML schemata settle on the usage of Akoma Ntoso in order to model and represent legal documents and also logic capabilities concerning the procedural rules that exists in such documents [11]. In the same spirit, another project related to ELI's framework is European Case Law Identifier (ECLI) which launched to model case laws [12]. In the context of the European project ESTRELLA, which established the Legal Knowledge Interchange Format(LKIF) MetaLex has been expanded [14].

Following the footsteps of other successful efforts in Europe aiming at modernizing the way Greek legislation is made public, a group of the Department of Informatics and Telecommunications at National and Kapodistrian University of Athens and the department

[4] https://en.wikipedia.org/wiki/Web_Ontology_Language
[5] https://en.wikipedia.org/wiki/Akoma_Ntoso

of Computer Science at University of Oxford at United Kingdom developed an ontology OWL, called Nomothesia [15].Ilias Chalkidis, together with the rest of his collaborators belonged to this group, after gathering a number of posted legal documents relying on the principles of ELI and MetaLex managed to translate almost 2676 legal files into 1.85 millions RDF using G3parser. Nomothesia G3 (Greek Government Gazette) parser is the first tool that introduced to populate Nomothesia API legislative data set. They utilize content records, which have been transformed from PDF files (distributed files of Government Gazette), as the leading accessible input, attempting to extricate legislative data (metadata, structure, legitimate content) and change it to RDF triples. Moreover, they injected a SPARQL endpoint and a RESTful API which enable the formulation of complex queries. Hence, all this effort and procedure resulted into a platform for querying Greek legislation.

3 Background

3.1Background on Diavgeia

3.1.1 Open Data

Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and share alike.

The most important precise details are:

- Availability and Access
- RE-use and Redistribution
- Universal Participation

If you're wondering why it is so important to be clear about what open means and why this definition is used, there's a simple answer: interoperability. Interoperability denotes the ability of diverse systems and organizations to work together (inter-operate). In this case, it is the ability to interoperate - or intermix - different datasets. Interoperability is important because it allows for different components to work together. This ability to componentize and to 'plug together' components is essential to building large, complex systems. Without interoperability this becomes near impossible — as evidenced in the most famous myth of the Tower of Babel where the (in) ability to communicate (to interoperate) resulted in the complete breakdown of the tower-building effort.

3.1.2 Diavgeia Portal

In a relevant effort to enforce the transparency of the state services, Greece enacts at 13/07/2010 a new law with title "Transparency enforcement: Posting of laws on internet – Updated Edition". The constituent aspect of this law is the admissibility of the compulsory posting of laws, government ordinances and acts that publish people on internet and the creation of conditions and procedures to ensure for the general exposure.

Therefore, in the general ambiance of uncertainty and suspicion towards the authorities of the Greek state, the clarity program was created with the aim of posting the laws as mentioned above to enhance transparency and established a new "social contract" between the citizen and the state

Simultaneously posted acts amplify institutionally by:

- Any act, despite those published in the Government Gazette; acquire ascendancy only by posting at Diavgeia. Upon completion, every act is digitally signed by Diavgeia system and a unique web number is accorded to it. Specifically, this number is the identity of each posted act certifies it.
- Every citizen can invoke the posted documents in his transactions with the public services without the need for their validation. It is enough to invoke the unique number of the act for their inherit search by the public services.
- The public services process cases without circulating the posted documents. It is enough to invoke the unique number of acts for the communication between the state services.

"Diavgeia" programme supports both the format of transparency used by Great Britain and an API. The citizen has the ability through the site to have access to various statistical data such as graphs that concern for example the unique number of acts that have been posted by category of the body or the number of transactions per type of transaction.

The main advantages of the program concern:

- Safeguarding transparency of government actions
- Eliminating corruption by exposing it more easily when it takes place
- Observing legacy and good administration
- Reinforcing citizen's constitutional rights, such as the participation in the Information System
- Enhancing and modernizing existing publication in formats that are easy to access, navigate acts and decisions

- Making of all administrative acts available in formats that are easy to access, navigate ad comprehend, regardless of the citizen’s knowledge level of the inner processes of the administration.

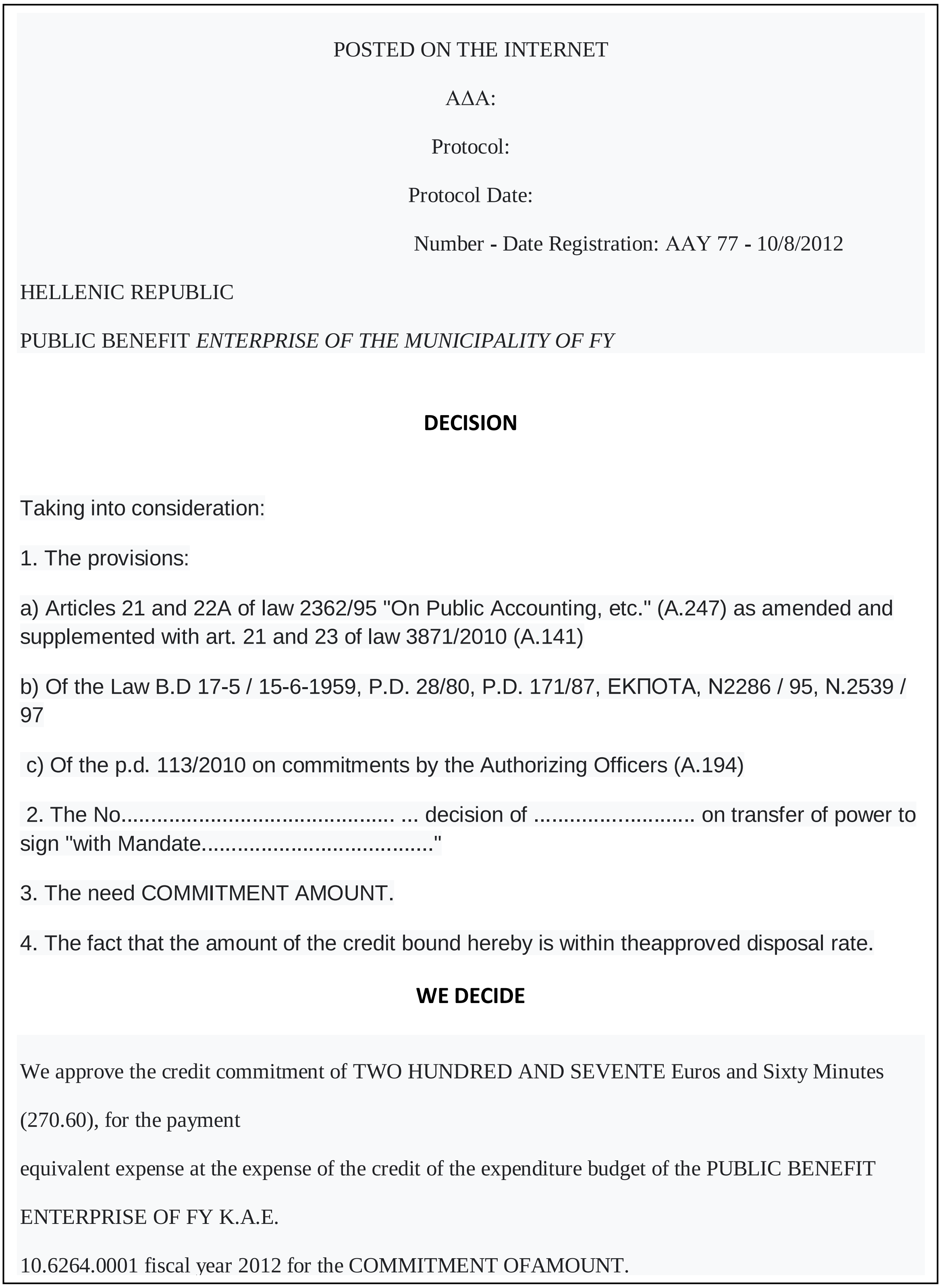
3.1.3 Metadata of Greek Legislation

As has already been reported, a vast amount of government decisions are released daily. These decisions cover a wide spectrum of operations in Greece. There are 34 different forms of decisions that can be uploaded to Diavgeia platform that have been adopted by the Greek government. Based on the background of the policy, the decisions type is selected by the government. I identify that most of them follow the same pattern, considering the very different decision type.

Legal documents are supplemented by metadata, in addition to the above procedural features. This fundamentally incorporates the title of the legitimate document, which must be detailed but brief so aw to reflect its substance, the type, the year of distribution, which can be gathered by the distribution date and the serial number. These last mentioned components (type, year, number) of metadata information serve as a specific identifier of the legal document. The topic and the sheet number of the government gazette in which the legal text is issued are both of similar significance.

When reference to other legislation or public administration decisions is necessary, this is done with citations. In order to be accurate and authenticated, quotes must contain the form of legal text, the number and the year of printing of the document.

In addition to the metadata information that comprises a legislative document, there is also a vast amount of secondary metadata material that may be added to a legal document. Various other elements that accompany a legislative text are elements that add information such as the signatories and their government position or specific geographical areas (locations).



POSTED ON THE INTERNET

ΑΔΑ:

Protocol:

Protocol Date:

Number - Date Registration: AAY 77 - 10/8/2012

HELLENIC REPUBLIC

PUBLIC BENEFIT *ENTERPRISE OF THE MUNICIPALITY OF FY*

DECISION

Taking into consideration:

1. The provisions:

a) Articles 21 and 22A of law 2362/95 "On Public Accounting, etc." (A.247) as amended and supplemented with art. 21 and 23 of law 3871/2010 (A.141)

b) Of the Law B.D 17-5 / 15-6-1959, P.D. 28/80, P.D. 171/87, ΕΚΠΟΤΑ, N2286 / 95, N.2539 / 97

c) Of the p.d. 113/2010 on commitments by the Authorizing Officers (A.194)

2. The No.. ... decision of .... on transfer of power to sign "with Mandate......................................"

3. The need COMMITMENT AMOUNT.

4. The fact that the amount of the credit bound hereby is within theapproved disposal rate.

WE DECIDE

We approve the credit commitment of TWO HUNDRED AND SEVENTE Euros and Sixty Minutes

(270.60), for the payment

equivalent expense at the expense of the credit of the expenditure budget of the PUBLIC BENEFIT

ENTERPRISE OF FY K.A.E.

10.6264.0001 fiscal year 2012 for the COMMITMENT OFAMOUNT.

Fig. 1. Article 1 of Presidential Decree 2012/54

3.2 Named Entity Recognition

Named-entity recognition (NER) (also known as (named) entity identification, entity chunking, and entity extraction) is a subtask of information extraction that seeks to locate and classify named entities mentioned in unstructured text into pre-defined categories such as person names, organizations, locations, medical codes, time expressions, quantities, monetary values, percentages, etc. A named entity is a term in data mining that specifically distinguishes one object from a collection of other items with similar attributes. Named entity delimits something that a proper name refers to. Persons, organizations, product names, quantities, monetary values, events for example. Other examples include proteins, genes, drugs, disorders from the biomedical domain.

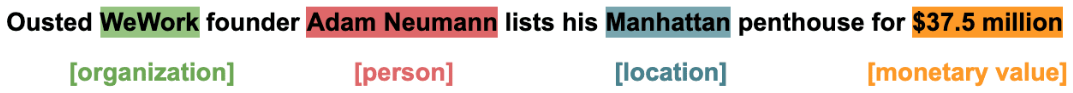

Ousted WeWork founder Adam Neumann lists his Manhattan penthouse for $37.5 million
[organization]
[person]
[location]
[monetary value]

Fig. 2. Named entity recognition task implementation

The identification and classification of named entities in text is one of the most common tasks in Natural Language Processing (NLP). The lack of broad data sets of languages other than English is among the problems encountered. This makes it difficult to use conventional machine learning approaches or neural networks to classify named entities in Greek texts.

The task of recognizing named entities first arisen at the 6th Message Understanding Conference (MUC-6). The conference's objective was to extract structured information, related to corporate and defense activities from unstructured text, including articles in newspaper. It was acclaimed that it is vital to define units of data, including time, date, money, percentage as well as, people, places, organizations and numerical expressions. These references are listed in the text as the name of Named Entity Recognition and Classification [16].

The conference introducing MUC was as it was the springboard for the persistent advancement of projects for the purpose of viable named entity recognition. Specifically the Information Retrieval and Extraction (IREX) project [17], Automatic Content Extraction

(ACE) program [16], Conference on Natural Language Learning 2002, 2003 (CoNLL 2002 and 2003) [19][20] contributed significantly for the clarification and research related to NER.

As mentioned above, named entity recognition in Greek language and especially in legislative documents is a difficult and demanding task due to lack of tools and the complexity of Greek language [21]. Academic projects relevant to Greek legal documents have taken place in the past few years. Koniaris [22] present their technique for the automatic development of a Greek legislative text representation of Akoma Ntoso. In addition, Beris and Koumparakis depict their strategy for re-engineering of Diavgeia resulting express the public decisions as Linked Open Data [23]. They offer an open source implementation, called DiavgeiaRedefined, which produces and visualizes decisions inside a Web Browser, provides a SPARQL endpoint to retrieve and query these decisions and enables citizens with an automated mechanism to review accuracy and identify potential foul play by an opponent.

One more step in the effort for a more effective named entity recognition regarding Greek Legislation documents was enacted from Chalkidis, Angelidis, Koubarakis using state-of-the-art deep neural networks and word embeddings. Word embeddings are pre-trained using unsupervised algorithms over large corpora based on the linguistic observation that similar words tend to co-occur in similar contexts (phrases).So, word embeddings capture both semantic and syntactic information as well as correlations between words. Evaluating the performance of three LSTM-based model for named entity recognition in legal documents, they manage to create a new vocabulary from the textual references they extracted from documents, link them with entities from other third party data sets and finally provide a new data set for Greek geo-landmarks [24].

Another approach for modeling and mining legal sources was Solon, a legal document management platform [25]. The DSL (Domain Specific Language) approach for parsing legal documents, identifying their structural elements and metadata, and converting plain text to XML following the LegalDocML schema [26].

3.3Knowledge graphs

A set of interlinked descriptions of entities-objects, events or concepts is described by a knowledge graph. Via linking and semantic metadata, knowledge graphs position data in context and thus provide a platform for data aggregation, unification, analytics and sharing where:

- Descriptions have formal semantics that allow efficient and unambiguous processing by both human and computers.
- The descriptions of the entities refer to each other, creating a network in which each entity represents a part of the description of entities linked to it and provides a context for their interpretation.

Generally KGs display improved capabilities to expose higher-order interdependencies in otherwise unstructured data. For several, information graphs (KGs) are important to NLP assignments, although it is time-consuming and costly to create a stable domain specific KG. A variety of methods have been proposed to build KGs with minimal human interference.[27][28]

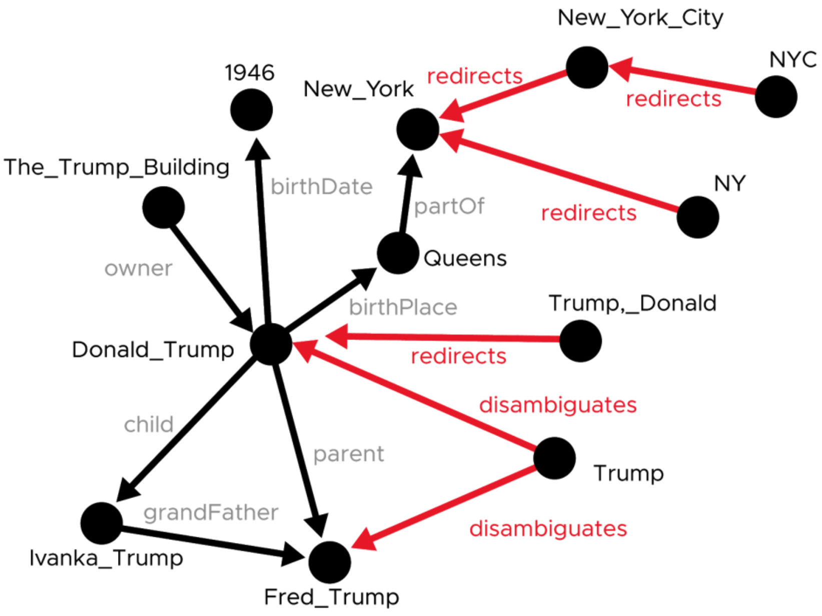


Fig. 3. Named Entity Recognition with Knowledge graph

4. Materials and Methods

In this chapter, we present and analyze the process we used to extract the data from Diavgeia portal, and methods we used for extracting structured information from the documents

The goal of this thesis is to extract structured information from public documents, available on the Diavgeia portal and subsequently gain insights by statistical analysis and visualizations. The whole processtakes place on Google's Colab which is an open-source cloud service based on Jupyter Notebook that supports free CPU.[6]

4.1 Data Model

Each act, document consists of the following parts:

- A set of metadata which describes the purpose and content of the act, as well as the issuer of the act (the carrier, the unit, the signatory).
- The document of the act in PDF format. The file is marked with the Internet Posting Number (ADA) that has been assigned in practice, and is digitally signed by Diavgeia system.
- An optional set of accompanying documents (attachments).
- A version number (Version ID).
- Finally, all posted laws are assigned a unique Internet Posting Number (ADA).

Specifically for the metadata of a legal document, they are divided into the following categories:

- Basically or common metadata. This kind of information which is available independently of all acts. Such information is the type of legal document, the subject, the protocol number, the date of issue of the document, the date of the last amendment, etc.

[6] https://colab.research.google.com/

- Special metadata. This kind of information whose existence depends on the type of the legal document. For each different type, an alternative set of specific fields is specified.

We present an example of the JSON structure that is returned when retrieving information from a posted act.

```json
{
  "ada": "BΛEZN-P7N",
  "protocolNumber": "A.200",
  "subject": "Transplantation Act",
  "issueDate": 1380585600000,
  "organizationId": "30",
  "signerIds": [
    "28006"
  ],
  "unitIds": [
    "10000"
  ],
  "decisionTypeId": "Γ.3.5",
  "thematicCategoryIds": [
    "611"
  ],
  "extraFieldValues": {
    "eidosYpMetavolis": "Transplantation",
    "documentType": "ACT",
    "relatedDecisions": [
      {
        "relatedDecisionsADA": "BΛ9ΛN-32N"
      }
    ]
  },
    "privateData": false,

    "submissionTimestamp": 1389700140183,
    "status": "PUBLISHED",
    "versionId": "6deb2804-ef9b-4640-8a4e-92f2acecb970",

    "documentChecksum": "034a544f359fe152ca646347502653c5276c786a",
    "attachments": [
        {
            "id": "c5e2ad78-6c6b-452a-8ef9-4cae2b9b691a",
            "description": "Cover file",
            "filename": "attachment.pdf",
            "mimeType": "application/pdf",
            "checksum": "e357f6261bcd123bb868535ddeb58471c3265ca7"
        }
    ]
}
```

Fig. 4. JSON file extracted from Diavgeia

4.2 Communication with Diavgeia API

Connecting with Diavgeia API is achieved using python language and a library called requests. This library is used to be able to use the properties of the HTTP protocol. Specifically, allows us to use URLs to provide information vie Web. The initial interface with Diavgeia portal is done using samples of encoders available on Githuband on Diavgeia platform. In the first step, we create an account on Diavgeia site. Next, we fill the necessary information in the sample code provided by the platform, in order to access the API. After an extensive research and browsing on Diavgeia portal and after readingseveral posted decisions we ended up extracting and processing a specific type of decision: “Commitment decision”.

4.3Analysis Methods

The dataset used in this thesis is created by extracting files from Diavgeia portal. Since there is not an available dataset the following pipeline is followed in order to generate a balanced dataset for the named entity recognition task. The pipeline includes the following steps which will be later further analyzed:

- Data Parsing
- Data set Preprocessing
- Entity Recognition

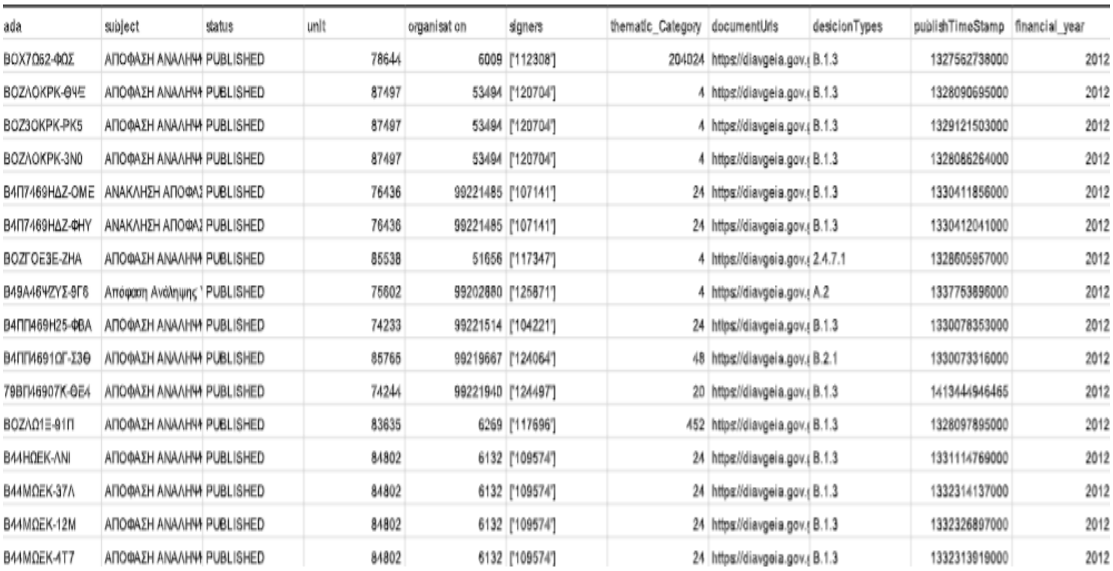
4.3.1 Data Parsing

The next step is to define our search format using the formats we explained earlier. In this case we use a URL of the form:

```
http://diavgeia.gov.gr/opendata/search/advanced.json?q=subject:"ΑΠΟΦΑΣΗ ΑΝΑΛΗΨΗΣ
ΥΠΟΧΡΕΩΣΗΣ"  AND issueDate:[DT(' + key + ')' + 'TO DT(' + val + ')]
```

, where the key and val elements are specified from a set of dictionaries. As already mentioned, Diavgeia portal has started to upload laws and decisions on the platform in 2010. However, we decided to extract files in order to analyze them starting from 2012 until 2019. For this reason, we prepare 8 different files following the construction and context of dictionary in python. Each file is representing one year and also it is separated by month. The purpose of creating these dictionary-files is to be able to pull all the laws and decisions that are posted each year but also to keep the date of issuance of each file. Based on the data we had at our disposal as offered to us by the archive provided by Diavgeia, we divided the data into lists according to their type and then these lists were converted into a table with the help of the Pandas library. The first representation of the imported data is available below:

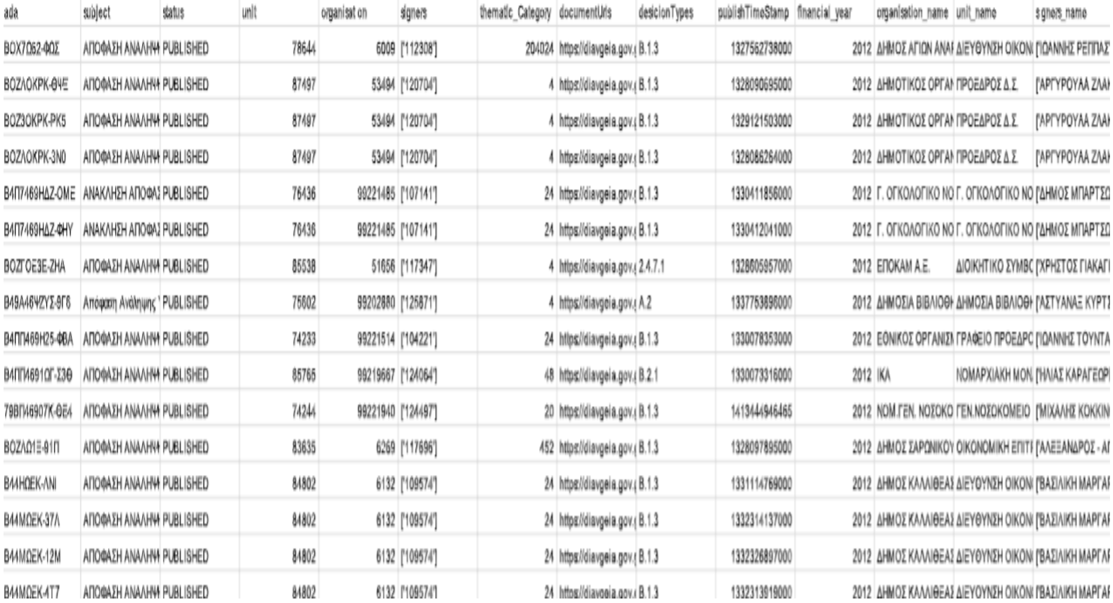
Table 1.Table with the first elements extracted



ada	subject	status	unit	organisat on	signers	thematic_Category	documentUrls	desicionTypes	publishTimeStamp	financial_year
ΒΟΧ7Ω62-ΦΩΣ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	78644	6009	['112308']	204024	https://diavgeia.gov.	B.1.3	1327562738000	2012
ΒΟΖΛΟΚΡΚ-Θ4Ξ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	87497	53494	['120704']	4	https://diavgeia.gov.	B.1.3	1328090695000	2012
ΒΟΖ3ΟΚΡΚ-ΡΚ5	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	87497	53494	['120704']	4	https://diavgeia.gov.	B.1.3	1329121503000	2012
ΒΟΖΛΟΚΡΚ-3Ν0	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	87497	53494	['120704']	4	https://diavgeia.gov.	B.1.3	1328086264000	2012
Β4Π7469ΗΔΖ-ΟΜΞ	ΑΝΑΚΛΗΣΗ ΑΠΟΦΑΣ	PUBLISHED	76436	99221485	['107141']	24	https://diavgeia.gov.	B.1.3	1330411856000	2012
Β4Π7469ΗΔΖ-ΦΗΥ	ΑΝΑΚΛΗΣΗ ΑΠΟΦΑΣ	PUBLISHED	76436	99221485	['107141']	24	https://diavgeia.gov.	B.1.3	1330412041000	2012
ΒΟΖΓΟΞ3Ε-ΖΗΑ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	85538	51656	['117347']	4	https://diavgeia.gov.	2.4.7.1	1328605957000	2012
Β49Α46ΨΖΥΣ-9Γ6	Απόφαση Ανάληψης '	PUBLISHED	75602	99202880	['125871']	4	https://diavgeia.gov.	A.2	1337753896000	2012
Β4ΠΠ469Η25-ΦΒΑ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	74233	99221514	['104221']	24	https://diavgeia.gov.	B.1.3	1330078353000	2012
Β4ΠΠΙ4691ΩΓ-Σ3Θ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	85765	99219667	['124064']	48	https://diavgeia.gov.	B.2.1	1330073316000	2012
79ΒΓΙ46907Κ-ΘΞ4	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	74244	99221940	['124497']	20	https://diavgeia.gov.	B.1.3	1413444946465	2012
ΒΟΖΛΩ1Ξ-91Π	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	83635	6269	['117696']	452	https://diavgeia.gov.	B.1.3	1328097895000	2012
Β44ΗΩΕΚ-ΛΝΙ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	B.1.3	1331114769000	2012
Β44ΜΩΕΚ-37Λ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	B.1.3	1332314137000	2012
Β44ΜΩΕΚ-12Μ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	B.1.3	1332326897000	2012
Β44ΜΩΕΚ-4Τ7	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	B.1.3	1332313919000	2012

The data set that created consists of 4000 decisions. However, as can be seen from the above illustration, the values under some columns such as the signatories and the organizational unit are not entirely clear to which entity they are referring to because their value is just an identity number, an id. For this reason we wanted to translate this unique number to which person or organization respectively belongs. Diavgeia portal provides files in the format of XML which contains a catalogue with the names of signatories for example in correspondence with the id which they refer to the JSON files are available. Therefore, after a time of consuming and demanding process we managed to match the ids with the corresponding names. Hence, the table that was finally created was more comprehensive and enriched regarding the metadata provided for each decision.

Table 2.Table with corresponding organization, unit and signatories names



ada	subject	status	unit	organisation	signers	thematic_Category	documentUrls	desicionTypes	publishTimeStamp	financial_year	organisation_name	unit_name	signers_name
ΒΟΧ7Ω62-ΦΩΣ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	78644	6009	['112308']	204024	https://diavgeia.gov.	Β.1.3	1327562738000	2012	ΔΗΜΟΣ ΑΓΙΩΝ ΑΝΑ	ΔΙΕΥΘΥΝΣΗ ΟΙΚΟΝ	['ΙΩΑΝΝΗΣ ΡΕΠΠΑΣ
ΒΟΖΛΟΚΡΚ-ΘΨΞ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	87497	53494	['120704']	4	https://diavgeia.gov.	Β.1.3	1328090695000	2012	ΔΗΜΟΤΙΚΟΣ ΟΡΓΑΝ	ΠΡΟΕΔΡΟΣ Δ.Σ.	['ΑΡΓΥΡΟΥΛΑ ΖΛΑ
ΒΟΖ3ΟΚΡΚ-ΡΚ5	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	87497	53494	['120704']	4	https://diavgeia.gov.	Β.1.3	1329121503000	2012	ΔΗΜΟΤΙΚΟΣ ΟΡΓΑΝ	ΠΡΟΕΔΡΟΣ Δ.Σ.	['ΑΡΓΥΡΟΥΛΑ ΖΛΑ
ΒΟΖΛΟΚΡΚ-3Ν0	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	87497	53494	['120704']	4	https://diavgeia.gov.	Β.1.3	1328086264000	2012	ΔΗΜΟΤΙΚΟΣ ΟΡΓΑΝ	ΠΡΟΕΔΡΟΣ Δ.Σ.	['ΑΡΓΥΡΟΥΛΑ ΖΛΑ
Β4Π7469ΗΔΖ-ΟΜΞ	ΑΝΑΚΛΗΣΗ ΑΠΟΦΑ	PUBLISHED	76436	99221485	['107141']	24	https://diavgeia.gov.	Β.1.3	1330411856000	2012	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟ	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟ	['ΔΗΜΟΣ ΜΠΑΡΤΣΩ
Β4Π7469ΗΔΖ-ΦΗΥ	ΑΝΑΚΛΗΣΗ ΑΠΟΦΑ	PUBLISHED	76436	99221485	['107141']	24	https://diavgeia.gov.	Β.1.3	1330412041000	2012	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟ	Γ. ΟΓΚΟΛΟΓΙΚΟ ΝΟ	['ΔΗΜΟΣ ΜΠΑΡΤΣΩ
ΒΟΖΓΟΞ3Ε-ΖΗΑ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	85538	51656	['117347']	4	https://diavgeia.gov.	2.4.7.1	1328605957000	2012	ΕΠΟΚΑΜ Α.Ε.	ΔΙΟΙΚΗΤΙΚΟ ΣΥΜΒΟ	['ΧΡΗΣΤΟΣ ΓΙΑΚΑΓ
Β49Α46ΨΖΥΣ-9Γ6	Απόφαση Ανάληψης '	PUBLISHED	75602	99202880	['125871']	4	https://diavgeia.gov.	Α.2	1337753896000	2012	ΔΗΜΟΣΙΑ ΒΙΒΛΙΟΘΗ	ΔΗΜΟΣΙΑ ΒΙΒΛΙΟΘΗ	['ΑΣΤΥΑΝΑΞ ΚΥΡΤ
Β4ΠΠ469Η25-ΦΒΑ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	74233	99221514	['104221']	24	https://diavgeia.gov.	Β.1.3	1330078353000	2012	ΕΘΝΙΚΟΣ ΟΡΓΑΝΙΣΜ	ΓΡΑΦΕΙΟ ΠΡΟΕΔΡΟ	['ΙΩΑΝΝΗΣ ΤΟΥΝΤΑ
Β4ΠΠ4691ΩΓ-Σ3Θ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	85765	99219667	['124064']	48	https://diavgeia.gov.	Β.2.1	1330073316000	2012	ΙΚΑ	ΝΟΜΑΡΧΙΑΚΗ ΜΟΝ	['ΗΛΙΑΣ ΚΑΡΑΓΕΩΡ
79ΒΓΙ46907Κ-ΘΞ4	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	74244	99221940	['124497']	20	https://diavgeia.gov.	Β.1.3	1413444946465	2012	ΝΟΜ.ΓΕΝ. ΝΟΣΟΚΟ	ΓΕΝ.ΝΟΣΟΚΟΜΕΙΟ	['ΜΙΧΑΛΗΣ ΚΟΚΚΙΝ
ΒΟΖΛΩ1Ξ-91Π	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	83635	6269	['117696']	452	https://diavgeia.gov.	Β.1.3	1328097895000	2012	ΔΗΜΟΣ ΣΑΡΩΝΙΚΟΥ	ΟΙΚΟΝΟΜΙΚΗ ΕΠΙΤ	['ΑΛΕΞΑΝΔΡΟΣ - Α
Β44ΗΩΕΚ-ΛΝΙ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	Β.1.3	1331114769000	2012	ΔΗΜΟΣ ΚΑΛΛΙΘΕΑΣ	ΔΙΕΥΘΥΝΣΗ ΟΙΚΟΝ	['ΒΑΣΙΛΙΚΗ ΜΑΡΓΑ
Β44ΜΩΕΚ-37Λ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	Β.1.3	1332314137000	2012	ΔΗΜΟΣ ΚΑΛΛΙΘΕΑΣ	ΔΙΕΥΘΥΝΣΗ ΟΙΚΟΝ	['ΒΑΣΙΛΙΚΗ ΜΑΡΓΑ
Β44ΜΩΕΚ-12Μ	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	Β.1.3	1332326897000	2012	ΔΗΜΟΣ ΚΑΛΛΙΘΕΑΣ	ΔΙΕΥΘΥΝΣΗ ΟΙΚΟΝ	['ΒΑΣΙΛΙΚΗ ΜΑΡΓΑ
Β44ΜΩΕΚ-4Τ7	ΑΠΟΦΑΣΗ ΑΝΑΛΗΨ	PUBLISHED	84802	6132	['109574']	24	https://diavgeia.gov.	Β.1.3	1332313919000	2012	ΔΗΜΟΣ ΚΑΛΛΙΘΕΑΣ	ΔΙΕΥΘΥΝΣΗ ΟΙΚΟΝ	['ΒΑΣΙΛΙΚΗ ΜΑΡΓΑ

The principal goal of our problem is to implement named entity recognition concerning the main text of the decisions. As mentioned above, the information extraction using Diavgeia API provides a JSON format file including the data and metadata which have been extensively presented. However, the only information that is available from the JSON file is a URL which leads to the basic content of the law. As a result, we created a python-function which downloads the file via the URL path and stores this file to my Google drive account from where we could upload them.

4.3.2 Data Preprocessing

Data scientists spend most of their time cleaning and preprocessing the data rather than modeling. Text is an unstructured form of data and could contain noisy content that could mislead our model, thus a basic text cleaning is a necessary step to build more robust models and also expect more reliable results. Legislation is an event-driven process. Judicial documents shall be passed on the basis of an effective parliamentary process and published in the Government Gazette. Later governing documents pay be changed with regard to their substance (alternations) or may be repealed. In the course of this process, we ought to catch the structure of the legal contract and the progression of its material over time. By the analysis of a large corpus of Greek legislation it seemed substantial to move on text preprocessing. We used the spacy, Natural Language Toolkit (NLTK) and Regular Expressions (RE) Python libraries for the implementation of data preprocessing. The techniques for text cleaning utilized for this problem are presented as follows:

Punctuation removal: Punctuation marks and non-alphabetic symbols, such as currency, do not add any significant information. Only dots are kept (.) as they act as the breaking point between sentences.

Tokenization: Tokenization is the process of splitting a sentence into a list of words (tokens). It is an essential part of the preprocessing phase as it converts unstructured text to a sequence of features. For example, the string "The food tastes so good" becomes "the", "food", "tastes", "so", "good".

There are two more techniques that are utilized for text cleaning but in our case we applied them only for specific features. Those are:

Stop-words removal: Stop-words are the words that are used very frequently and they somewhat lose their semantic meaning. Words like "ένα","είναι", "αυτά","και" , etc. are some examples of stop-words. The approach used to remove the stop-words was to have a predetermined list and filter them out from the tokenized sentences.

After the extensive searching and reading over the available decisions, we found that some words appeared in all the decisions. Hence, we decided to add to the already existing stop words as well. For example, the word 'ΑΔΑ' appears in all files as we mentioned since it is the distinctive number that every decision has. However due to the fact that the next step of our research is called Named Entity Recognition is based on the format of the words and mainly on whether the first letter or the whole word are uppercase, we did not proceed to convert all words into lowercase or uppercase. In addition, a specific word that belongs to stop words but in lowercase, 'την' was added to stop words with the first letter being capitalized.

4.3.3 Entity Recognition

The final step in our research is the Named Entity Recognition (NER).Named entity recognition is one of the simplest of the common message understanding tasks. The objective is to identify and categorize all members of certain categories of "proper names" from a given corpus. In the wide spectrum of NLP tools there are a plenty of different tools for NER. However, their functionality and their primary highlight is that they are trained within the English language. In spite of this, there are a few tools available that they have either have been made exclusively for Greek language or at least Greek language along with others have been included as an additional feature and functionality in existing ones. In our quest to identify the tool that will have the best and most effective results in terms of entity recognition, we applied several tools to our data set. Some of them are:

- Polyglot NER: A system that builds annotators for 40 major languages using Wikipedia and Freebase. This approach, which introduced by Rami Al-Rfou on 2014, does not require NER human annotated datasets or language specific resources like tree banks, parallel corpora and orthographic rules[28].
- Natural Language Toolkit NER: NLTK provides a classifier that has already been trained to recognize named entities, accessed with the function nltk.ne_chunk (). The classifier adds category labels such as PERSON, ORGANIZATION, and GPE.
- Spacy: Spacy is free open source library for Natural Language Processing in Python. It features NER, Pos tagging, dependency parsing, word vectors and more. Spacy tool is implemented for the usage by many languages. Named entity annotations were created by Giannis Daras using Prodigy and the OntoNotes 5 annotation schema [30].

Eventually we settled on Spacy tool for the task of entity recognition. For this purpose, the “el_core_web_lg” pretrained model is utilized on Colab environment.

4.4.4 Graph implementation

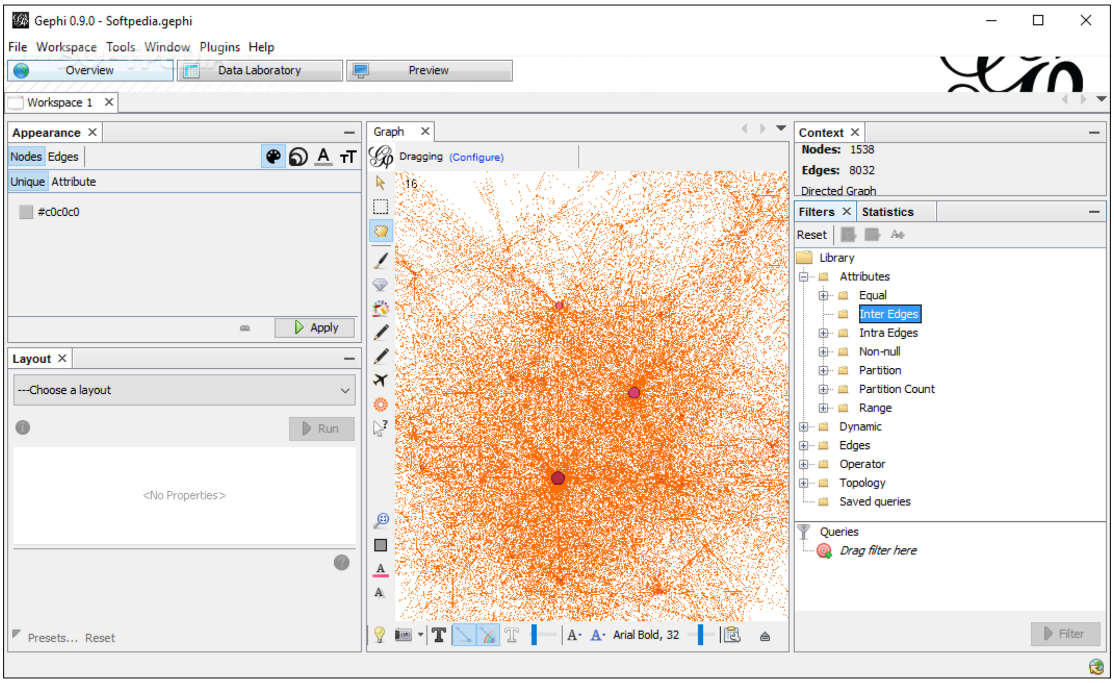
The last step of the research is the implementation of a knowledge graph. A graph is a non-linear data structure consisting of nodes and edges. The nodes are sometimes also referred to as vertices and the edges are lines connect any two nodes in the graph. The knowledgegraph represents a collection of interlinked descriptions of entities – objects, events or concepts. Knowledgegraphs put data in context via linking and semantic metadata and this way provide a framework for data integration, unification, analytics and sharing. In the case of our research and the graph that we are going to construct, we set as nodes the extracted entities from the legal documents. In addition, we consider that the entities extracted from the same document are related to each other. Hence, we combine these entities in all possible pairs for each file, for example a collection of 4 entities can form 6 different combinations.Finally, we measure the weights of the edges between the nodes. As weights we declare the occurrence of each pair across the total amount of documents.For this purpose we implement the python library Networkx and the Gephi and we create an undirected graph using edge weights.

4.4.4.1 Networkx

Networkx is a python language program for the discovery and study of the network and network algorithms. The main package includes the data framework for the representation of several types of networks or graphs, including simple graphs, directed graphs, and parallel and self-loop graphs. Networkx does not provide an interface, although one of its strengths is the ability to link existing code and libraries in a natural way that enables the incorporation of several resources [31].

4.4.4.2 Gephi

Gephi is an open source graph and network analysis program. Using a 3D render engine to present massive networks in real-time and speed up discovery. Flexible and multi-task architecture provides new ways to interact with diverse data sets and deliver useful visual outputs [32]. The interface (figure 3) is designed to workspaces where different and separated analysis can be performed. The user can quickly apply an algorithm, filter or tool to the software with little programming knowledge. Sets of nodes and edges may be collected either manually or by means of file system.



Nodes: 1538
Edges: 8032

Fig. 5. A screenshot of Gephi beta version 0.9.2

3.4.5 Network measures

Networks can be analyzed in several different ways. One way is to analyze their features, such as size, density, topology and statistical properties.

The most basic structural properties are the size and the density of a network. These properties are conceptually similar to the mass and composition of matter—they just tell us how much stuff is in it, but they do not inform us anything about how the matter is organized initially. Nonetheless, they are still the most fundamental characteristics, which are particularly important when comparing multiple networks. In this section very common graph measures are presented: the size, the density, the centralities, the modularity, the clustering coefficient and the transitivity. Since our analysis involves directed networks, with and without weights, the measures presented below are suited to directed networks only.

Structure of network

- *Size*

  The size of a network is characterized by the number of nodes and edges in i
- *Density*

  The density of a network is the fraction between 0 and 1 that tells us what portion of all possible edges is actually realized in the network. Foe a network G made of n nodes and m edges, the density p (G) is given by

  $$p(G) = \frac{m}{\frac{n(n-1)}{2}} = \frac{2}{n(n-1)}$$

  for an undirected network, or

  $$p(G) = \frac{m}{n-1}$$
- *Network Diameter*

  The diameter of a graph is the maximum eccentricity of any vertex in the graph. That is, it is the greatest distance between any pair of vertices. To find the diameter of a graph, it is necessary firstly to find the shortest path between each pair of vertices. The greatest length of any of these paths is the diameter of the graph.

Centralities

Centrality analysis decides the significance of vertices in an organize based on their network inside the organize structure. It may be a broadly utilized method to examine network-structured information.

- ➢ *Degree* A self arrange of the vertices of a graph can be built up by sorting them concurring to their degree. The comparing centrality degree degree-centrality (Cd) is characterized as Cd(v) := deg(v) [33].

- *Betweenness* Betweenness centrality of a node is the probability for the shortest path between two random chosen nodes to go through that node.In a weighted network the links connecting the nodes are no longer treated as binary interactions, but are weighted in proportion to their capacity, influence, frequency, etc., which adds another dimension of heterogeneity within the network beyond the topological effects. A node's strength in a weighted network is given by the sum of the weights of its adjacent edges.

$$\boldsymbol{s_i} = \sum_{j=1}^{N} a_{ij}\ w_{ij}$$

  With a_{ij} and w_{ij} being adjacency and weight matrices between nodes i and j , respectively.

- *Closeness*It is an inverse of the average distance from node i to all other nodes.

$$C(x) = \frac{N}{\sum_y d(y,x)}$$

  Where N is the number of nodes in the graph and d(y,x) is the distance between vertices x and y.

- *Eigenvector* Eigenvector centrality measures the “importance” of each node by considering each incoming edge to the node an “endorsement” from its neighbor. This differs from degree centrality because, in the calculation of eigenvector centrality, endorsements coming from more important nodes count as more. Another completely different, but mathematically equivalent, interpretation of eigenvector centrality is that it counts the number of walks from any node in the network that reach node i in t steps, with t taken to infinity.

Modularity

Modularity is a measure of the structure of graphs or networks. It is designed to measure the ability to split a network into modules or groups. Large networks have multiple connections between nodes of other nodes within their group; however there are loose connections between nodes of other groups. It is often used as an optimization method for detecting structures on social networks. However, it has been shown to be unable to identify very small communities.

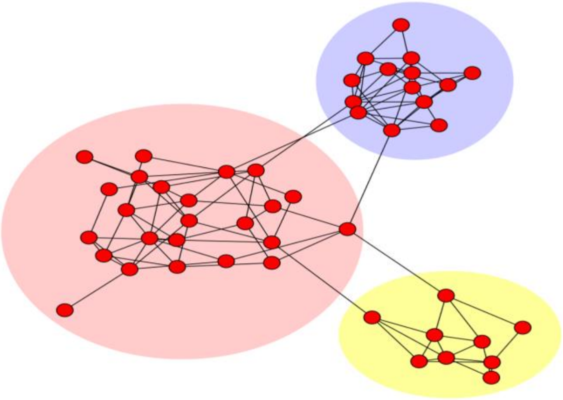


Fig 6. Modularity on networks

Clustering coefficient

The clustering coefficient of a graph is based on a local clustering coefficient for each node

$$C_i = \frac{number\ of\ triangles\ connected\ to\ node\ i}{number\ of\ triples\ centered\ around\ node\ i}$$

where a triple centered around node i is a set of two edges connected to node i.

The clustering coefficient for the whole graph is the average of the local values Ci

$$C = \frac{1}{n}\sum_{i=1}^{n} C_i$$

where n is the number of nodes in the network.

Transitivity

The transitivity T of a graph is based on the relative number of triangles in the graph, compared to total number of connected triples of nodes.

$$T = \frac{3\ x\ number\ of\ trianges\ in\ the\ network}{number\ of\ connected\ triples\ of\ nides\ in\ the\ network}$$

The transitivity of a graph is closely related to the clustering coefficient of a graph, as both measure the relative frequency of triangles.

5 Experimental Results

This section presents the experimental results for extracting structured information from documents extracted from the Diavgeia portal and subsequently providing better analyses and visualization on the extracted structured information Moreover,we present our implementation of the above mentionedmeasures in the python software environment and the graph visualization and analysis platform Gephi. We create two different data sets, one with weights between the nodes and one without weights. Then, we store the data frames in CSV format to drive. In addition, we have at our disposal the data frame from the metadata provided by each file extracted from Diavgeia combined with the assignment of the id in which some attributes are expressed such as organization, unit and signers.First, we illustrate some main statistical results concerning the amount of documents in terms of the signers and the organization of each legal document. Secondly, we present the results from the statistical analysis of the first data set which is formed to anundirected graph considering edges' weights and then we will display a comparison between the two data sets.

5.1 Data set analysis

The first approach of the analysis of the gathered data is to group some information included in the data frame and interpret them.

Table 3. Top 5 signers and organizations in terms of number of documents

Id	Name	Number of documents
	Signers	
106223	Themistoklis Kouimtzis	534
110182	Asterios Zografos	272
100002253	Xristos Tsirogiannis	137
100021483	Dimitrios Kranis	127
124788	Ioannis Rodis	109
	Organization	
99201081	Administration Body for Delta Loydia-Aksiou-Aliakmona	534
6451	Municipality of Polygyros	272
99221989	General Hospital 'Asklipeio' Voulas	146
6045	Municipality of Arta	137
99200399	National School of Judges (N.S.JU.)	129

The above table displays that for the period 2012-2019 the person who appears to have signed the most decisions is Themistoklis Kouimtzis who is mentioned as the signer to 534 documents. Moreover, the part of the table which presents the top five organizations in terms of their presence in the legal files from Diavgeia illustrates that "Administration Body for Delta Loydia-Aksiou-Aliakmona" seems to be the organizationwith which there were most financial transactions since the type of decisions we select to go over is the "Commitment decisions".

5.2Networkx and graph fashioning

One powerful python library which is used extensively for graph implementation is Networkx. Networkx is a python package for the creation, manipulation, and study of the structure, dynamics, and functions of a complex networks. Graphs in networkx can be grown in several ways since it includes many graph generator functions and facilities to read and write graphs in many formats. For the purpose of this thesis we select the undirected graph. Below we present the basic properties after the creation of the undirected graph.

Info graph

Number of nodes: 4781

Number of edges: 25703

Average degree: 10.839

5.3 Graph analysis

5.3.1 Graph visualization

Secondly, we continue analyzing the data in Gephi by importing the two files mentioned above. We create a project and we proceed to the data laboratory tab, we select the edges tab and then the import spreadsheet option. Finally, in the pop up window that appears we choose the spreader in which the csv file is located and we select as separator "comma", as table "Edges table" and as char set the option "UTF-8" in order Gephi to identify Greek characters. In the next tab we have to pay attention and check the option "Create missing nodes" to create the nodes of the graph. Returning to the Overview tab, our graph is now visible. Before starting using the filters and statistics offered by Gephi, we will use the layouts provided to get a better picture of our network. As we mentioned in the basic information of the network that we have already implement in python environment the graph consists of 4781 nodes and 25688 edges. The layout we apply with this number of nodes is the Force Atlas 2. Using the layout for several minutes our graph acquires a form which does not change anymore, at this point we stop the algorithm and notice that several small clusters have formed. At this point it is substantial to mention that despite the clusters that formed

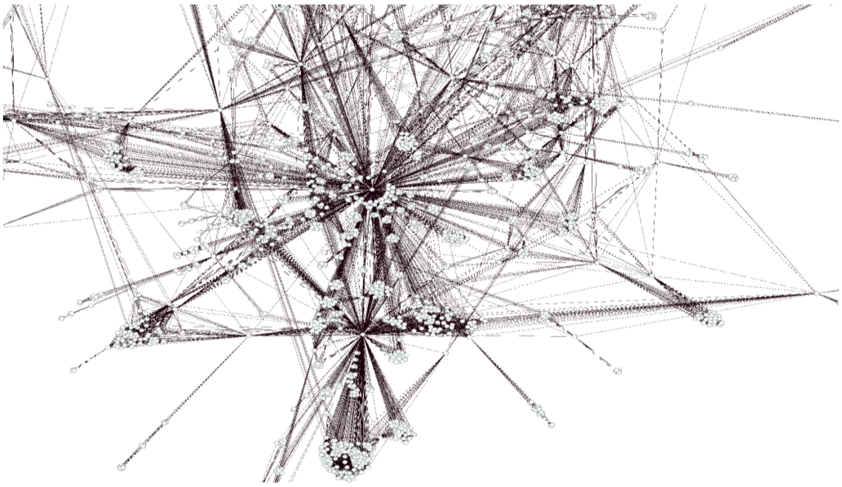


Fig. 7.Snapshoot of the network implemented in Gephi

Regardless of the clusters that have been formed, there are modes either individually or in small clusters of two, three or four nodes which are detached from the main graph. This is not obvious in the graph above because we decided to present an enlarged image in order the basic structure of the network to be clear.

5.3.2 Statistical analysis

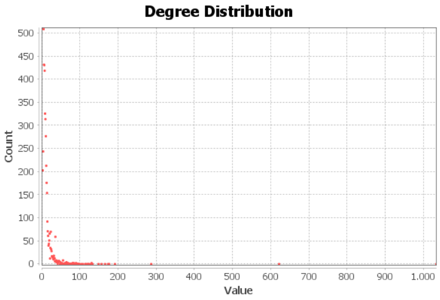
Next step of our graph analysis is to use the various statistics provided by Gephi. These network measures are: Average Degree, Average Weighted Degree, Net diameter, Graph Density, Modularity, Average Clustering Coefficient, Average Path Length and Transitivity. The parameters which are will set to these metrics are displayed below:

- Modularity: We select *Randomize*, check the option *use weights* (for weighted graph and uncheck for un weighted graph) and also set the variable *Resolution* to 1.0.
- The layout Force Atlas 2: We select *Prevent Overlap*.

The results are available in both graph tab compiling for the whole graph but also in the data laboratory where the above statistical data are listed in detail for each node.

Table4.Overview of statistical analysis of the graph

Network Measures	**Directed graph with weights**
Average Degree	10.839
Average Weighted Degree	19.72
Net Diameter	10
Graph Density	0.002
Modularity	0.858
Avg. Clustering coefficient	0.915
Avg. Path Length	3.744
Transitivity	0.200



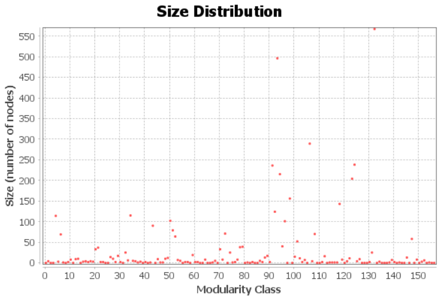


Fig. 8 . Degree distribution   Fig. 9 . Modularity Class Distribution

Connected Components

Gephi permits to calculate the number of nodes that are not associated at all or are weakly connected with other nodes. As we have already described the structure of the graph, we mentioned that there are nodes which are not located at the core of network. The exact number of these components which are weakly connected is 119.

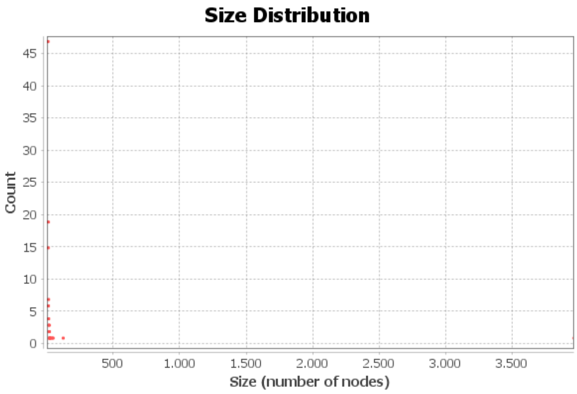


Fig. 10. Number of weakly connected components

Centralities

Centrality measures are a vital tool for understanding systems, regularly too known as graphs. These calculations utilize chart hypothesis to calculate the significance of any given node in a network. They cut through noisy data, uncovering parts of the organize that require consideration – but they all work in an unexpected way. Each degree has its own definition of “importance”.In our case of our research we apply named entity recognition in the decisions and we extract a plenty of entities which in turn formed the amount of nodes for our graph. We perform 3 different centrality measures: Betweenness, Closeness and Eigenvector. Using networkx for this purpose, this procedure results in a list of tuples where each tuple consists of the node and the corresponding degree of centrality in descending order. Hence, we choose to continue analyzing the top 10 nodes based on each centrality measure.

Table 5.Top 10 nodes concerning Betweenness centrality

	Nodes	**Value**	**Label**
Betweenness	EE	0.345	ORG
	EE EE	0.134	ORG
	IKA	0.098	ORG
	Decide-Approve	0.081	ORG
	Διά	0.072	ORG
	ΠΟΥ	0.054	ORG
	ΚΑΕ	0.038	ORG
	ΥΠΟΙΚ	0.035	ORG
	New Architecture of Decentralized Administration Kallikratis FEK	0.032	ORG
	ΔΣ	0.030	ORG

Table 6.Top 10 nodes concerning Closeness centrality

	Nodes	**Value**	**Label**
Closeness	EE	0.391	ORG
	EE EE	0.339	ORG
	Διά	0.330	ORG
	EURO	0.320	ORG
	ΚΑΕ	0.318	ORG
	ΓΛΚ	0.314	ORG
	Βεβ	0.313	ORG
	ΠΟΥ	0.311	ORG
	ΕΣΠΑ	0.311	ORG
	PRESIDENT	0.310	PERSON

Table 7.Top 10 nodes concerning Eigenvector centrality

	Nodes	Value	Label
Eigenvector	EE	0.479	ORG
	EE EE	0.331	ORG
	Development Fund Hellenic Republic Ministry of Culture Culture of Sports Museum Natural History of the Petrified Forest of Lesvos Sigri Lesvos	0.099	ORG
	Information center of Mitilini	0.099	ORG
	Museum Natural History of Petrified Forest of Lesvos	0.094	ORG
	Organization for the construction of the New Museum of Acropolis	0.084	ORG
	Εγκεκρ Προυπ	0.074	ORG
	ΚΔ	0.073	ORG
	Intellectual Property Related Rights Cultural Issues	0.071	ORG
	ΕΛΚΕ ΕΜΠ	0.070	ORG

Observing the results from the three centrality degrees, we can notice that same nodes appear in all cases but the order in which they appear changes. For example, the nodes " ΠΟΥ" , "ΚΑΕ" and "Διά". The node "ΠΟΥ" is takes the 5th position regarding Betweenness centrality and the 7th concerning the Closeness. However, obviously the first two position and by extension the two nodes that gather most of the nodes are "EE" and "EEEE".

Modularity

Another statistical measure which is connected with centrality and specifically with Betweenness centrality is modularity. We have already analyzed the definition of graph modularity and in addition we presented the calculated modularity of our graph implemented in Gephi interface. Gephi provide us the ability to edit the graph that we have created and append additional features aiming to the visualization of how the nodes are grouped. So, we proceed to the Overview tab again and in the appearance option we select nodes and then the tab Ranking. There is a dropdown selection list with the statistical measures applied at the graph. We choose the measure Modularity class and we push the Apply button. The following figure illustrates the result from the procedure which is described previously.

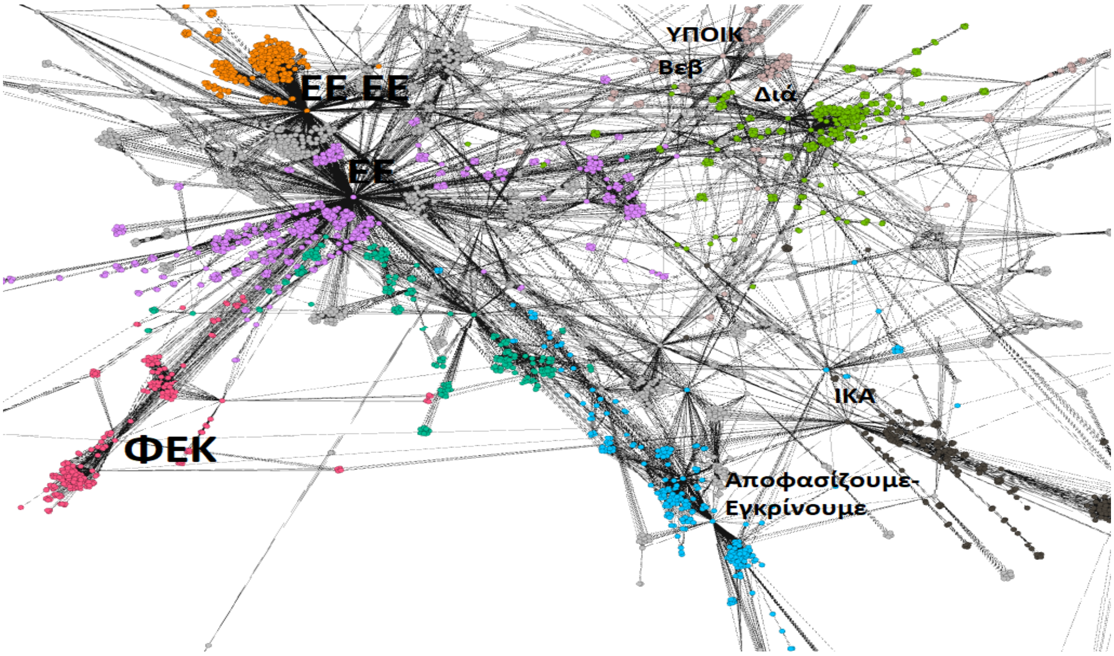


Fig 11. Modularity Implementation at Gephi

Comparison with graph without weights

The graph we produce and analyze is an undirected graph including weights. After researching and working with Gephi tool we identify that during the implementation of most network measures, there is an option *use edge weights.* This finding enables us to find out whether a graph with weights or without weights provides differences in the statistical study of a network. For this reason we repeat the process of importing the csv file into Gephi but this time we do not change the “Weight” column departing the weight for all edges to value 1. We recalculate all the network measures and below we display our findings.

Table 8.Results from graph comparison

	Undirected graph with weights	**Undirected graph without weights**
Average Degree	10.839	10.839
Average Weighted Degree	19.72	10.985
Net Diameter	10	10
Graph Density	0.002	0.002
Modularity	0.856	0.870
Avg. Clustering coefficient	0.915	0.481
Avg. Path Length	3.744	4.973

It is indicated that for both graph implementations all of the metrics provide the same outcomes except for modularity and Avg. weighted degree.

6 Discussion

In the previous section, we present the results from the extraction of structured information for Diavgeia portal. Due to the fact that the main task was the Named entity recognition on the legal documents we extracted, we concluded that for the purposes of this task the best results were provided by Spacy. We conclude that the more important entities in term of centrality are “ΕΕ” and “ΕΕ ΕΕ”. Despite the limitations which will be presented below, we created a knowledge graph by which we collect information in terms of the dispersion of the entities-nodes. However, although Spacy provided us the best results it terms of the Named Entity Recognition task in contrast with other available tools the outcomes of the entity recognition were not effective. There were a plenty of extracted entities where whether they were extracted incorrectly or they assigned with wrong label.

Despite, this partial ineffectiveness at the results using this tool, we gain substantial insights from the information extraction and the graph implementation. Gephi provided us the ability to create real time graphs, to analyze them with statistical measures in order to explore more deeply the structure of a network. Moreover, using networkx gave us equally significant features regarding the analysis of the graph such as the calculation of the centralities, the most important most and the most frequent signers and organizations which are mentioned at the legal files we extracted from Diavgeia portal.

However, we aware that our research may have some limitations. First, in our experiments, we used a relatively small dataset including legal documents based on a specific category. Second, one of the most basic and important limitation was the manipulation of the Greek language. Since the majority of state-of-the-art tools used for Natural Language Processing (NLP) have been built and trained in English Language, we could not find and apply an existing tool for effective Named Entity Recognition on Greek. More specifically, we experimented with existing statistical machine learning models for named entity recognition on the Greek text in our work. We did not find any deep learning implementation of Greek named entity recognition in Diavgeia, which is suggested as a future research direction. As deep learning has shown state-of-the-art results in many NLP tasks, there is a need for a more customized trained model for named entity recognition in Diavgeia. Despite the fact that Spacy provides a pre-trained model based on Greek language, the specific structure of the texts of Greek legislation redounded to the efficiency of the process. This resulted in the task of entity recognition to receive outcomes like “ΠΟΥ” or “Διά” or “Αποφασίζουμε-Εγκρίνουμε” which faulty recognized as entitiesand alsoassigned with label ORG. However, the limited existence of NLP tool for named entity recognition based on Greek language is only one of the reasons for the results mentioned above.

Another factor that contributed is the format in which documents regarding Greek legislation are published. This factor contributed to the difficulty we encountered during the preprocessing process. We applied most of the main preprocessing methods on extracted documents like punctuation removal, text tokenization, stop words removal. Especially for the stop words removal there were some words which in some documents appeared in lowercase, in others the first letter or the whole word in uppercase. Therefore, although these works were present in the amount of stop words (in lowercase format) during the removal process they didn’t fend off. To avoid this limitation, the extensive analysis of decisions could end up to a vocabulary where words or even phrases that appear to the majority of the documents would be interpreted to attributes which are successfully recognized by an entity recognizer.

7 Conclusions and Future Work

In this thesis, we present the results from the extraction of structured information for Diavgeia portal. In addition, we display research results for previous approaches regarding the attempt to transform legal documents to public open data and also to extract structured information from these documents Our aim is the collection of documents from Diavgeia and the extraction of structured information from raw text to increase the data quality and subsequently include better capabilities for data analysis and visualization towards better transparency to the users.

First we connected with the API of Diavgeia portal and we extracted legal documents of the category “Commitment Decision” from 2012 until 2019. Working on Colab environment and using pandas library we formed a data frame containing information of the metadata provided from the JSON file we received. Moreover, we managed to interpret andmatch some metadata which were an id to the actual names, for example the signers and the organizations. Due to the fact that the main task was the Named entity recognition on the legal documents we extracted, we concluded that for the purposes of this task the best results provided by Spacy. We extracted in total 4781 entities. The challenge and the main object of our research were to extract information from legal texts that can add an additional value and also to be used for visualization purposes. The next step was to construct a knowledge graph from those entities in order to network that these entities formed. Using the python library Networkx and the Gephi tool we managed to create an undirected graph with 4781 nodes and 25.703 edges. Visualizing the network we built, we had a clearer picture of how the nodes are distributed. Finally, we identified the most important nodes based on centrality measures and also we displayed statistical analysis for the graph.

Since the dataset used for our experiments is relatively small (4000legal documents), a possible future direction is to test the Spacy library and the task of named entity recognition with larger datasets. There is a need for a labeled dataset with named entities in Diavgeia for training more accurate machine learning models. The labeling could be accomplished through manual annotation by public servants or through semi-automated approaches like active learning and distant supervision. The key through active learning is that a machine learning calculation can accomplish more prominent exactness with less labeled preparing occurrences in the event that it is permitted to select the preparing information from which is learns [34]. Distant supervision is the process of specifying the concept which the individual words of a passage, usually a sentence, are trying to convey. A potential improvement is to apply neural network model for effective preprocessing and named entity recognition.

7 Bibliography

[1]Colpaert P., Joye S., M. P. , MannersE. and Van de Walle R. (2013) The 5 stars of openportals. Proceedings MeTTeG 2013(7) 61-67

[2]Hoekstra, R.. "The MetaLex Document Server - Legal Documents as Versioned Linked Data." *International Semantic Web Conference* (2011).

[3] Boer, Alexander & Hoekstra, Rinke &Winkels, Radboud. (2002). MetaLex: Legislation in XML.

[4] Frosterus M., Tuominen J., Wahlroos M., Hyvönen E. (2013) The Finnish Law as a Linked Data Service. In: Cimiano P., Fernández M., Lopez V., Schlobach S., Völker J. (eds) The Semantic Web: ESWC 2013 Satellite Events. ESWC 2013. Lecture Notes in Computer Science, vol 7955. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-41242-4_46I

[5] Francesconi E., Kuster M. W, Gratz P., Thalen S., (2015) The ontology-Based Approach of the publications Office of the EU for Document Accessibility and Open Data Services International Conference on electronic Government and the Information Systems Perspective DOI:10.1007/978-3-319-22389-6_3

[6] G. Barabucci, L. Cervone, M. Palmirani, S. Peroni, and F. Vitali. Multi-layer Markup and Ontological Structures in Akoma Ntoso. In Int. Workshops AICOL-I/IVR-XXIV on AI Approaches to the Complexity of Legal Systems { ComplexSystems, the Semantic Web, Ontologies, Argumentation, and Dialogue, 2009.

[7] A. Boer, R. Hoekstra, R. Winkels, T. van Engers, and F. Willaert. METAlex: Legislation in XML. In JURIX: The Fifteenth Annual Conference, London, 2002.

[8] A. Boer, R. Winkels, T. van Engers, and E. de Maat. Time and versions in METAlex XML. In Proceeding of the Workshop on Legislative XML, Kobaek Strand, 2004

[9] P. Casanovas, M. Palmirani, S. Peroni, T. M. van Engers, and F. Vitali. Semantic Web for the Legal Domain: The next step. Semantic Web, 7(3):213{227, 2016.

[10] ELI Task Force. ELI - A technical implementation guide, 2015.

[11] M. Palmirani and F. Vitali. Akoma-Ntoso for legal documents. In Legislative XMLfor the Semantic Web. 2011.

[12] T. Athan, G. Governatori, M. Palmirani, A. Paschke, and A. Z. Wyner. Legalruleml: Design principles and foundations. In Reasoning Web. Web Logic Rules, 2015

[13] . M. V. Opijnen. European Case Law Identi_er: Indispensable Asset for Legal Information Retrieval. In From Information to Knowledge, volume 236 of Frontiersin Artificial Intelligence and Applications. 2011.

[14] M. Palmirani and F. Vitali. Akoma-Ntoso for legal documents. In Legislative XMLfor the Semantic Web. 2011.

[15] Chalkidis I., Nikolaou C., Soursos P., Koubarakis M. (2017) Modeling and Querying Greek Legislation Using Semantic Web Technologies. In: Blomqvist E., Maynard D., Gangemi A., Hoekstra R., Hitzler P., Hartig O. (eds) The Semantic Web. ESWC 2017. Lecture Notes in Computer Science, vol 10249. Springer, Cham. https://doi.org/10.1007/978-3-319-58068-5_36

[16]R. Grishman and B. Sundheim, \Message Understanding Conference-6: A Brief History," in Proceedings of the 16th Conference on Computational Linguistics -Volume 1, COLING '96, (Stroudsburg, PA, USA), pp. 466{471, Association for Computational Linguistics, 1996

[17] S. Sekine and H. Isahara, \IREX: IR & IE Evaluation Project in Japanese," inProceedings of the Second International Conference on Language Resources andEvaluation, LREC 2000, 31 May - June 2, 2000, Athens, Greece, European LanguageResources Association, 2000.

[18] G. R. Doddington, A. Mitchell, M. A. Przybocki, L. A. Ramshaw, S. M. Strassel,and R. M. Weischedel, The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation," in Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004,Lisbon, Portugal, European Language Resources Association, 2004.

[19] E. F. Tjong Kim Sang, Introduction to the CoNLL-2002 Shared Task:Language Independent Named Entity Recognition," in Proceedings of the 6th Conference on Natural Language Learning - Volume 20, COLING-02, (Stroudsburg, PA, USA), pp. 1{4, Association for Computational Linguistics, 2002.

[20] E. F. T. K. Sang and F. D. Meulder, \Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition," CoRR, vol. cs.CL/0306050, 2003.

[21] Garofalakis J., PlessasK., Plessas A. and Spiliotopoulou P. A project for the transformation of greek legal documents into legal open data Proceedings of the 22nd Pan-Hellenic Conference on Informatics, (144-149)

[22]Koniaris M., Papastefanatos G. and Vasilioy G. Towards Automatic Structuring and Semantic Indexing of Legal Documents Proceedings of the 20th Pan-Hellenic Conference on Informatics (2016) Article No.:4 Pages 1-6

[23] Beris T., Koubarakis M. (2018) Modeling and Preserving Greek Government Decisions Using Semantic Web Technologies and Permissionless Block chains. In: Gangemi A. et al.

(eds)The Semantic Web. ESWC 2018. Lecture Notes in Computer Science, vol 10843. Springer, Cham. https://doi.org/10.1007/978-3-319-93417-4_6

[24] Angelidis, I. et al. “Named Entity Recognition and Generation for Greek Legislation,” JURIX (2018)

[25] Koniaris M., Papastefanatos G., Meimaris M. and Alexiou G. 2017. Introducing Solon: A Semantic Platform for Managing Legal Sources Research and Advanced Technology for Digital Libraries. 10.1007/978-3-319-67008-9_53, (603-607)

[26] Koniaris, M.; Papastefanatos, G.; Anagnostopoulos, I. Solon: A Holistic Approach for Modelling, Managing and Mining Legal Sources. *Algorithms* **2018**, *11*, 196.

[27] Yu S., He, T. % Class, J. (2020) Constructing a Knowledge Graph from Unstructured Documents without External Alignment. ArXiv, abs/2008.08995

[28] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. Ernie: Enhanced language representation with informative entities. In Proceedings of the 57th AnnualMeeting of the Association for Computational Linguistics, pages 1441–1451.

[29] POLYGLOT-NER: Massive Multilingual Named Entity Recognition Rami Al-Rfou, Bryan Perozzi, Steven Skiena, Vivek Kulkarni, October 2014 DOI: 10.1137/1.9781611974010.66

[30] Daras Giannis, Adding Greek Language to spacy (2018), Github repository, https://giannisdaras.github.io/talks/

[31] Hagberg, A. et al. “Exploring Network Structure, Dynamics, and Function using Networkx.” (2008)

[32] Bastian, M. et al. “Gephi: An Open Source for Exploring and Manipulating Networks.” ICWSM (2009)

[33] Dwyer, T. et al. “Visual analysis of network centralities.” APVIS (2006)

[34]Settles, B. “Active Learning Literature Survey.” (2009)

[35] Lathrop, D. and Ruma, L., 2010. *Open government: Collaboration, transparency, and participation in practice*. " O'Reilly Media, Inc.".