Map4Scrutiny - A Linked Open Data Solution For Politicians Interest Registers

Inês Lopes

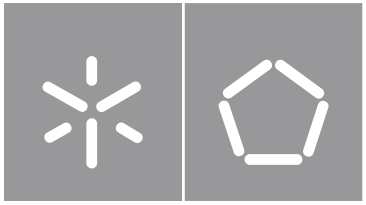
UMinho | 2021

Universidade do Minho
Escola de Engenharia

Inês Catarina Barreira Lopes

Map4Scrutiny - A Linked Open Data Solution For Politicians Interest Registers

December 2021



Universidade do Minho
Escola de Engenharia

Inês Catarina Barreira Lopes

Map4Scrutiny – A Linked Open Data Solution For Politicians Interest Registers

Masters Dissertation
Master's in Information Systems

Work developed under the supervision of:
Professor Doctor Ana Alice Baptista
Professor Doctor Óscar João Atanázio Afonso

December 2021

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration.

I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

RESUMO

O trabalho desenvolvido no âmbito desta dissertação descreve o processo de recolha, uniformização e transformação de dados abertos em formato de texto e tabelas (CSV) para dados abertos ligados (*Linked Open Data*). Especificamente, dados sobre os registos de interesses dos deputados à assembleia da república portuguesa e contratação pública, ligados pelas organizações que são mencionadas em ambos.
O estado da arte inclui uma análise de fundo aos conceitos de corrupção, transparência, dados abertos, e dados abertos ligados, tal como a projetos de dados abertos e dados abertos ligados relevantes.
A seleção dos dados a utilizar, com respeito aos tópicos de conjuntos de dados relevantes e ao interesse público, o desenho da solução proposta e a seleção de ferramentas, métodos e processos, seguiu a proposta de três ciclos de Hevner para uma abordagem ao desenho de investigação na ciência.
O processo de implementação é iniciado com a recolha de dados das fontes utilizando bibliotecas *Python* para *web Scraping* e a transformação dos mesmos em tabelas (CSV). Estes dados são depois limpos e uniformizados com auxílio do OpenRefine. Esta ferramenta é também usada para mapear os dados da tabela para triples que são exportados em ficheiros *Turtle*.
Este mapeamento foi previamente desenhado num perfil de aplicação que serviu de base para a criação das formas dos dados (ShExC) usadas para conduzir o processo de validação nos ficheiros *Turtle*. Esta validação assegura que os ficheiros gerados pelo OpenRefine são conformes com o perfil de aplicação.
Para descrever adequadamente os dados foram usados vocabulários já existentes complementados, quando necessário, com a criação de novas classes, propriedades e valores. Este processo está também descrito e os vocabulários estão disponíveis para consulta e reutilização.
Por fim, foram feitas consultas modelo em SPARQL para ilustrar a diferença entre os dados originais e o conjunto de dados transformado. O objetivo deste trabalho é contribuir para as áreas de dados abertos ligados e dados abertos para a transparência e escrutínio público. Os contributos principais para o primeiro são um novo esquema de dados e a descrição de todos os passos do processo de transformação. Para o segundo o contributo que se destaca é mais uma implementação que demonstra o potencial do escrutínio de dados no aumento da transparência através da comparação entra as consultas possíveis aos conjuntos de dados originais e ao resultante da solução proposta. O processo de implementação está documentado abaixo e os ficheiros resultantes disponibilizados para consulta.

Palavras-chave: Dados Governamentais Abertos, *Design Science Research*, Escrutínio, *Linked Open Data*, Transparência

ABSTRACT

The work developed in the scope of this dissertation describes the process of sourcing, uniformizing, and transforming text and tabular (CSV) open data to linked open data. More exactly, data on Portuguese parliamentarians' interest registers and public procurement, linked by the organisations mentioned in both.

The state of the art presented includes a background analysis on the concepts of corruption, transparency, open data, and linked open data and an analysis of relevant open data and linked open data projects.

The research was conducted using Hevner's three-cycle design science research approach which led to the definition of the data scope concerning relevant dataset topics and the public's interest, the design of the proposed solution, and the selected tools, methods, and processes.

The implementation process starts with Scraping the data from the sources with the aid of python libraries and generating tabular (CSV) outputs. These are cleaned and uniformized in OpenRefine. OpenRefine is also the tool used to map the data on the tables into triples and generate outputs in Turtle.

The map was designed in an application profile that also served as a base for writing the shapes (in ShExC) and conducting validation on the exported Turtle files. This validation ensures that the data is conformant with the application profile. To successfully describe the data in triples, on top of the external vocabularies used, new classes, properties and values had to be created. This process is also thoroughly described, and the outputs are open to access and reuse. Finally, sample SPARQL queries were made to showcase the difference between the sourced data and the resulting dataset.

The goal is to contribute to the field of linked open data and open data for transparency and public scrutiny. The main contributions to the first are a new data scheme and the description of every step in the transformation process, while to the latter the contribution is a further implementation showcasing the scrutiny potential of data in improving transparency by comparing the querying possibilities of the final dataset with the originals. Every step taken is documented below and the resulting outputs of the different stages are available for consultation.

Keywords: Design Science Research, Linked Open Data, Open Government Data, Scrutiny, Transparency

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

List of Annexes

List of Attachments

List of Acronyms

API	Application programming interface
CPI	Corruption Perception Index
CPV	common procurement vocabulary
CSV	Comma-separated values
DCTAP	Dublin Core Tabular Application Profile
DCMI	Dublin Core Metadata Initiative
DRE	Electronic official gazette
ER	Entity Relational Model
ESCO	European Occupations Taxonomy
EU	European Union
FOAF	Friend Of a Friend
GDPR	General Data Protection Regulation
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
ICT	Information and communications technology
ID	Identification
INE	National Institute of Statistics
INESC TEC	Institute for Systems and Computer Engineering, Technology and Science
IRI	Internationalized Resource Identifier
ISO	International Standards Organization
JSON	JavaScript Object Notation
LD	Linked data
LOD	Linked Open Data
LOV	Linked Open Vocabularies
NIF	Unique Fiscal Identifier
OCD	Ontologia Camera dei Deputati
OD	Open Data
ODB	Open data Barometer
ODC	Open Data Charter
OECD	Organisation for Economic Co-operation and Development
OGD	Open Government Data
OWL	Web Ontology Language
PDF	Portable Document Format
RDF	Resource Description Framework
RDFS	RDF Schema
ShEx	Shape Expressions
ShExC	Shape Expressions Compact
SKOS	Simple Knowledge Organization System
SPARQL	Simple Protocol and Rdf Query Language
SQL	Structured Query Language
TI	Transparency International
TTL	Turtle
URI	Unique Resource Identifier
URL	Uniform Resource Locator

WB	World Bank
WBCC	World Bank Control of Corruption
WWW	World Wide Web
WWWF	World Wide Web Foundation
XLSX	Excel Microsoft Office Open XML Format Spreadsheet file
XML	Extensible Markup Language

1 INTRODUCTION

Corruption is simultaneously a widely discussed subject and a misunderstood one. Storytelling in this area is highly focused on big scandals that tend to fade away undermining their importance and impact. These scandals are commonly triggered by data leaks. State-of-the-art research on corruption, Linked Open Data (LOD), and the connection between them returned few results for being a very specific and limiting connection. The scope of the search was then broadened to corruption and LOD separately and explored the individual concepts.

Back in 2014, the European Union (EU) issued a report exploring Open Data (OD) as a tool to fight corruption. This work portraits different projects across the EU advocating for the potential of data reuse as a corruption-fighting tool. These projects mostly feed on OD, re-organize and re-shape it to outputs more appealing and valuable to the stakeholders. [1]

Between government and institutional OD, and newspaper stories a lot of dispersed information is available for public scrutiny. However, the quality of said information, measured according to Tim Berners Lee's five star deployment scheme for OD [2], is often one-star OD or two stars at best. The full potentially of OD cannot be fully achieved unless the data is also linked and readable both for human usability and semantic web standards [3], this would be five-star quality OD or LOD. [2]

According to the LinkedData.center, to create LOD "Data can be extracted from different sources such as texts and databases and then represented as linked data using the RDF data model (...) relationships can be used to express connections between datasets as well as define concepts used in a dataset." [4]

Being this a broad idea, to achieve it involves several steps of previous work that need to be done to execute the work properly and achieve the best result.

The motivations for this research are the high Portuguese corruption perception index [5], public interest on the subject, and the wide range of available but poorly organized information. In addition, there is the previous work on OD as a corruption prevention tool [1] and understanding the added value of having it as five stars OD (LOD) [2] that also motivates the work developed below.

Most of the Portuguese government's data found on topics susceptible to corruption is not linked open data. Instead, the data is spread across multiple stand-alone datasets each with its own design and it is neither machine-readable nor interoperable by semantic web standards. In particular, the datasets on Parliamentarians Interest Registers and Public Procurement, two datasets, from two different sources that fit the criteria for increasing transparency and treating topics susceptible to corruption.

That constitutes the research problem at hand: The data opened by the government on Parliamentarians Interest Registers and Public Procurement is isolated, difficult to link, and neither machine-readable nor interoperable by semantic web standards. Thus, it is not ready to be used and reused by third parties.

Therefore, the main objective of this research is to propose a solution for linking open data on Portuguese parliamentarians interest registers and public procurement, transforming said data from their current level (one or two) to level five, LOD [2]. To achieve this objective, the datasets may be transformed to a machine-readable format, the two datasets may be linked to each other and external LOD vocabularies, and the final combined dataset may be uploaded into a Triplestore with an online SPARQL endpoint so it can be queried by anyone. Also, the process of transformation should be transparent so that any third party could replicate the proposed solution.

The research method used is Hevner's Three Cycle Design Science Approach [6]. The developed solution attempts to respond to the stated problem by basing the design decisions on known literature and previous projects.

This work is developed in the domain of information systems, more specifically semantic web and LOD with the area of the application being transparency. Identifying the area of application as transparency is more accurate than corruption because further work needs to be developed to understand the impacts of LOD implementations as corruption-fighting tools. The prototype and results presented here are promising but not enough.

The identified stakeholders are on first level journalists, corruption-fighting organizations, and researchers, and on a second level the general public.

The expected contributions include state-of-the-art research comprising a background analysis in corruption, open data and LOD illustrated with sample projects. A definition of the methodologies, a description of the implementation from the sourcing of the data to the process of mapping it to RDF-Turtle and publication in a Triplestore. The files resulting from each phase are also shared and include: LOD Application Profile, code written to source the data, logs from the cleaning data process, transformed RDF-Turtle data, ontology, and value vocabularies, the shapes of the data in ShExC, the shape maps, the validation results, and sample SPARQL queries. Each of these outputs is also a contribution to the knowledge base in LOD.

The sample queries will illustrate the potential of an LOD profile to apply scrutiny on the relationships between persons and organizations of interest to the public. The ability to easily apply scrutiny to entities is useful when there is suspicion of corruption.

The remainder of this document is structured as follows: The State of the art, chapter two, is divided into Corruption, Open Data, and Linked Open Data. The first defines the area of application and explores transparency as a corruption-fighting tool, and open data or even LOD as transparency enablers. Open Data is discussed in the second sub-chapter with further analysis on the concepts of what makes good open data and what datasets are important. Throughout, open data availability, quality, and abiding laws are largely discussed with a special focus on Portugal. This information is then put against already ongoing projects focused on open data re-use to select national data sources. The third part goes into the domain of this work by exploring LOD in the scope defined by the previous two sub-chapters and the overall good practices and implementation methods.

Chapter three, Data And Methods, established the methods used to conduct the research and feeds on the previous chapter to set the data scope of the proposed solution.

Then in Implementation , every step taken in sourcing, cleaning and uniformizing, designing the LOD application profile, writing the shapes, selecting, and creating vocabularies, transformation, validation and publishing the data is described and generated outputs presented.

At last, chapter 5, Conclusions consolidates the work developed, compares the objectives with the results, and presents recommendations and future work.

2 STATE OF THE ART

The goal of the state of the art is to understand where the research stands in terms of perception of corruption, ways to fight it, the role of open data in this scenario and more specifically the role of LOD. Being this a dissertation, the first searches are conducted in academic paper databases looking for research work. When this first approach is not fruitful enough the search is complemented with reports and other online sources created by credible organizations such as the European Union, Transparency International, and The World Bank.

The search for similar projects in the area begins with an academic perspective by refining a query revolving around the keywords in the base idea: ontology, corruption, and LOD Table 1. Transparency is used as an alternative to corruption because it is a way to prevent it. According to TI, it is the duty to act visibly, predictably, and understandably to promote participation and accountability and allow third parties to easily perceive what actions are being performed. It is related to corruption in the sense that "institutions should make a commitment to report annually on the measures they are adopting to strengthen risk management, especially in relation to bribery and corruption. " [7]

For every document with an adequate title and abstract the list of references and citations is also viewed and adequate documents are selected. Most of these documents are not as helpful as one would expect. There is vast and relevant work on both ontologies and LOD, and then in corruption and public scrutiny. The scarcity begins to be noted when these concepts are mixed Table 1.

Table 1: The number of publications found in Scopus for the filtered queries.

Search Keywords	Domains	Number of Results	Source
Ontology AND Corruption	Computer Science, Engineering	24	Scopus
	Computer Science (all available)	1	Web Of Science
("linked data" OR "linked open data" OR "open data") AND (corruption OR transparency)	Computer Science, Engineering	578	Scopus
	Computer Science (all available), Information Science Library Systems, Public Administration	368	Web Of Science
("linked data" OR "linked open data") AND (corruption OR transparency)	Computer Science, Engineering	156	Scopus
	Computer Science (All), Information Science Library Systems, Public Administration	70	Web Of Science
("linked data" OR "linked open data") AND (corruption)	Computer Science, Engineering	12	Scopus
	Computer Science (All), Information Science Library Systems, Public Administration	7	Web Of Science

Since the results are limited the next step is to broaden the search. This led to a search on OD as a corruption-fighting tool. The main sources for understanding this subject are the EU, the Organisation for Economic Co-operation, and Development (OECD), the World Wide Web Foundation (WWWF), Transparency International (TI), and The World Bank (WB). The following section funnels down from a broad context of corruption to the Portuguese context followed by symptoms, and good practices.

2.1 Background Analysis - Corruption

Table 2 Definitions of Corruption as in [8]

Corruption Description	Primary Source
"Corruption is the abuse of entrusted power for private gain. It hurts everyone who depends on the integrity of people in a position of authority" [1], [9]	Transparency International
"The extent to which public power is exercised for private gain, including both petty and grand forms of corruption, as well as "capture" of the state by elites and private interests" [8], [10]	The World Bank Description
"Abuse of power for private gain." [8], [11]	European Commission
"Active corruption or 'active bribery' is defined as paying or promising to pay a bribe (OECD, 2008). The 'active or passive misuse of the powers of Public officials (appointed or elected) for private financial or other benefits" [8]	OECD
"Requesting, offering, giving or accepting, directly or indirectly, a bribe or any other undue advantage or prospect thereof, which distorts the proper performance of any duty or behaviour required of the recipient of the bribe, the undue advantage or the prospect thereof." [8]	Council of Europe

All in all, every description involves a form of using power for private gain, with a focus on TI's addition of "entrusted power", this is important when considering that corruption is a practice here applied to those who hold a public office and were in some way elected or entrusted with that position. [12]

The definitions are alike and seem to point in the same direction, however, there are some caveats. The concept of harming the public is ambiguous and dependant on context since there is no guideline to know what is in the interest of the public. [8] Also, corrupt practices operate behind the curtains, except for when cases are exposed [12]. This is an obstacle to any measure that tries to access corruption on its extent, impact, or perception as is seen further on.

Nevertheless, corruption was still conceptualized, and, according to [11] there are several known types defined depending on the actor. Grand corruption is applied when the actors hold high offices, while petty corruption refers to small bribes given, for instance to doctors for better treatment or examiners for a driving license. Corruption is also classified as local and central separating the range of the actions, such as in the municipality (local) or on a national level (central).

As seen above bribery is often mentioned meaning that the private gain is of the financial sort. Corrupt practices are estimated to annually cost the European economy 120 billion euros, out of these, 5.33 billion euros are associated with public procurement alone [12]. The latter is one of the most observed types of corruption. However, the financial cost alone is not enough of a measure, the practices mess with the social tissue and one of the most prominent and extreme connections with it is society's lack of trust in public institutions [12], [11]. Growing untrust becomes a problem in itself with consequences such as tax avoidance by discouraged taxpayers and even possibly a threat to democracy. [12]

2.1.1 European and Portuguese Contexts

Corruption measurements are either based on subjective data, such as the perception of corruption, or objective data such as judicial prosecutions. [8] All the indicators mentioned below fit the first category. The WB and Corruption Perception Index (CPI) are the two most widely used corruption indicators [8]. In common they have the fact that both present their results ordered by country ranks, creating a direct comparison between countries.

In 2014 a study developed by Malito comparing measures of corruption criticized both for different reasons: CPI questionably displays the perception of corruption as being connected to the extent of corruption, while the WB displays a biased perspective from large business elites. There was expressed concern over the fact that both aggregate many sources, risking the loss of clarity. It's important to keep these limitations in mind alongside the fact that more questions were raised than answered by this report. [8] In Table 3 they are both described alongside two Eurobarometer's developed by the European Commission.

Table 3 Measures of Corruption

Surveys and Indexes	**Institution**	**Goal**
Control of Corruption	The World Bank (2019)	It is one of the six worldwide governance indicators used by WB. In 2019 it accessed 209 countries using on average 9 sources per country ranging from 16 to only 1. It reflects the use of public power for private gain [13] and aims to aid in instrumentalizing government assistance. [8]
Corruption Perception Index	Transparency International (2019)	Focused on how corruption is seen and perceived in 180 countries around the world. It feeds from 13 sources from 12 institutions and only considers countries appearing on at least three of them. [5] The aim is to raise awareness and demand accountability from political leaders. [8]
Eurobarometer on businesses' attitudes to corruption	European Commission (2015)	A flash survey to EU countries focusing on their perception of corruption for companies when doing business. Focuses on bribery and public procurement management and overall corrupt procedures. [12]
Eurobarometer on corruption	European Commission (2014)	A survey was carried in 2013 for the commission's report on corruption in 2014. It was meant to be the first of a bi-annual practice, but it ended up being the only one.[11] They surveyed 27 786 persons on their perception and experience with the matter.

The definition from TI introduces the concept of accountability which is very often mentioned when discussing matters related to corruption. Accountability in the context of a business, or entity is the willingness to be judged on actions taken. To be accountable for something is to own upon what is in cause and accept consequences. [14]

Across the EU only 23% of citizens consider their governments' anti-corruption efforts to be effective, while 26% consider being personally affected by corruption in their daily lives, and 67% advocate for more

transparency and oversight over the financing of political parties [11]. Bribery and the use of connections are often considered the easiest way to get certain public services by 73 % of the respondents. [11] Companies also struggle against corruption, 56% say corruption in public procurement is widespread on a national level and 60% on a regional/local authority level. In Portugal, this number goes up to 68%. [11] Transparency in election processes and campaign donations are associated with low levels of corruption. Countries, where campaign finance regulations are comprehensive and systematically enforced, have an average CPI score of 70. Moreover, enforcing better practices in this matter was part of the policies in 60% of the countries that raised their CPI between 2012 and 2019. [5]

Although the data used by CPI is subjective, corruption scandals and prosecutions (absolute data) have been shown to directly influence the respective countries scores.

Table 4 Corruption Indexes Scores

Geographical Area	**CPI (2019)**		**WB - Control of Corruption (2019)**	
	Rank (1-180)	**Score (100-0)**	**Rank (0–100)**	**Score (-2.5–2.5)**
EU - Average	34	64	77.2	0.95
Portugal	30	62	77.4	0.76
Highest EU Score	Denmark - 1	87	Finland 2.15	99.04
Lowest EU Score	Bulgaria - 74	43	Bulgaria -0.16	50.43

The results presented in Table 4 place Portugal shortly below the EU average score for both CPI and WBCC (World Bank Control of Corruption). WBCC is only one of six indicators used by the WB, from these, the one where Portugal performed worst was Regulatory Quality which “Reflects perceptions of the ability of the government to formulate and implement sound policies and regulations that permit and promote private sector development.” [13] These scores, alongside information about businesses' high perception of corruption, raise the question of whether there is a lack of trust in the government? And if there is, what is causing it? There is not enough data to answer these questions now, but further research will keep them on the table.

2.1.2 Syndromes of Corruption

Some wrongdoings are more common than others, or at least appear so based on the corruption indicators results and reviews. Businesses often struggle with public procurement mainly in four fields: 1) Tailor-made specifications for certain bidders; 2) Splitting of public tenders into smaller bids to seem less attractive to competitors; 3) Prominence of conflicts of interest; 4) Excessive use of emergency claims to avoid publication. [12]

Deep-rooted corruption is an obstacle to economic development. However, the "Mobility of labour between the public and private sectors are essential for the functioning of modern society and can bring major benefits to both the public and the private sector." [11] For a healthy relationship between the private and public sectors and a balance between both statements above, the risk of public officials' giving preferential treatment to former or current private-sector connections must be mitigated.[11]

2.1.3 Good Practices

Before listing good practices, one should be aware that corruption-fighting measures are highly dependent on social, political, and legislative contexts. [11] Policies efficient in one country might not be on its neighbour if there are social nuances, and this is especially true when it comes to law and third-party agencies, as is explained further down. The impact of legislation and administrative measures is dependent on the cultural mentality of both public bodies and society. To sum it up, recommendations fit contexts, not countries. [12]

One of the suggested ways to improve cultural mentality is by further involving the public in political decisions while diminishing the influence of money. Political consultation needs to be fair, impartial, and non-partisan. [5]

A key pillar of the fight against corruption is prevention. [12] If, as seen in the previous chapter, public officials economic interests are a risk factor, then there are two options: either to prohibit certain activities deemed susceptible to corruptive practices, establishing "cooling-off" periods for an efficient and verified transfer, and applying sanctions; Or to increase accountability and transparency through an effective asset and interests declaration, making possible transgressions easier to spot and possibly corruptive practices accessible to public scrutiny. This also improves public trust. [12], [11]

Both [12], [11] advert for thorough verification of asset declarations, preferably done by third party agencies provided with enough power, tools, and access to conduct investigation efficiently and spot the lack of or incorrect information. When this verification is involved in red tape, or it is not sufficiently

endorsed it can be a problem. Verification really is the key here. When assets declarations exist but lack verification, they can create the illusion of transparency and accountability while not fully doing their part. Many member states lack these tools, but a good example is Romania where a third-party agency is capable of scrutinizing and validating more than half a million declarations a year. This work has led to finding both omitted and wrong information and caused public officials to lose their offices or seats. Preventive policies have had great success when implemented effectively, and not so much when it is done in fragments.

2.1.3.1 Open and Transparent

"Transparency is a particularly strong tool in the fight against corruption. Freedom of access to information improves good governance and helps to make government more accountable." [12]

Above, trust in the government is linked to corruption levels, preventive measures are deemed mandatory, and transparency seems to be linked to both. The remaining good practices were therefore narrowed down to those focusing on transparency and openness in a preventive way. The more transparency there is in the public sector, the fewer opportunities for corruption there are, making markets more competitive in a healthy way. For these practices to be transparent the processes, results, and documents must be open.

Preventive measures are showcasing public contracting and expenditures in at least Croatia, Estonia, Portugal and Slovenia [12]. Portugal's platform is Base.gov.pt where all public contracts are uploaded by the hiring company. "The publication of contracts in both BASE and the Official Gazette is now mandatory for direct adjustments, increases of 15 % in the price of already concluded contracts and potential penalties." [11] To this day, the fact that data is not time-sensitive, and contracts being published with months delay is still a limitation.

High standards of transparency in public procurement are a recommendation of the EU as of 2014. It is included in the preventive policies alongside easy access to public interest information and the assertion that it helps mitigate corruption in public administration and political party financing.

Achieving transparency using OD became a trend in such a way that 21 EU members are part of the Open Government Partnership whose goals are to promote transparency, empower citizens, fight corruption, and harness new technologies to strengthen governance.

There is no simple guideline that works in every situation. However, reductions in corruption have been seen in countries lacking regulation and strategic programmes but with a preventive and tactical high transparency standards approach. The same cannot be said for the implementation of complex legal and institutional programs in countries with serious corruption problems. [12], [11]

2.1.4 Final considerations

2.1.4.2 Agencies and government programs

These complex programs and the act of anti-corruption agencies are not a lost cause, they both work but are highly dependent on context. Agencies run the risk of creating the appearance that corruption is being dealt with but not being given the tools to do this job. Simultaneously, agencies achieved amazing results when their powers and legitimacy were not under constant pressure, and they have the cooperation of public entities.[11]

Strategies and programs are more complicated since they have, in some cases, been essential and in others had little effect. The objective reasons why they succeed, or fail are not explained.

2.1.4.3 Legislation

To conclude this chapter, and thus turn solely to data, it is essential to look at some legislation. "An effective policy response needs to be based on evidence about its prevalence and forms in a given country, the conditions that enable it and the institutional and other incentives that can be used against it." [12]

Studies have found a positive link between the quality of the legal system and the level of institutional trust. [15] Also, good data is needed to hold governments accountable for their actions [16].

When the rules are not effectively enforced systemic problems remain unattended, and there are no concrete results. "Track records of prosecutions and final court decisions are weak, and few cases of public procurement corruption are finalised with dissuasive sanctions." [11]

Formal cooperation is not enough if people remain under pressure to be silent. Corruption networks are often dismantled through whistle-blowers, so there needs to be a culture aided and supported by the law to protect these people.

The United Kingdom (UK) and Ireland are great examples of legislation about whistleblowing. There, whistle-blowers are protected and encouraged to make any wrongdoing public no matter the intentions behind the revelations or the employment status. There is also financial aid if whistleblowing is the cause of dismissal. Another example is the Netherlands, one of the few member states where integrity policies are successfully an active part of public administration both locally and nationally. [12]

For law enforcement to work, it must be both independent and equipped to fight corruption. "Likewise, striking the right balance between privileges and immunities of the public officials and ensuring that these are not used as obstacles to effective investigation and prosecution of corruption allegations is still an issue in some Member States." A workshop organised by the Commission in 2015 showed that law is more likely to be efficient when it comes out of public discussions. [12] Public interest and awareness is, once more, key for the success of law enforcement.

2.2 Background Analysis – Open Data

For the requirements of this project, all data sources used must be OD in the sense that it is available for consultation, but not necessarily including the possibility of bulk download. This chapter further explores ideas introduced in 2.1 Background Analysis - Corruption.

The 2.1.3 Good Practices sub-chapter lightly introduces the combination of transparency, corruption fight, and new technologies to strengthen governance and the importance of citizens' trust in their governments. Following this line, the OD mentioned in this chapter mostly comes from public sources, is often called open government data (OGD), and is tightly connected with e-government guidelines and goals. This makes the government the main provider of data of the public interest.

According to the Open Data Charter (ODC), "Open data is digital data that is made available with the technical and legal characteristics necessary for it to be freely used, reused, and redistributed by anyone, anytime, anywhere." [17] This can be further specified with G20's six principles for data: Open by Default; Timely and Comprehensive; Accessible and Usable; Comparable and Interoperable; Boosting Governance and Citizen Engagement; Aiding Inclusive Development and Innovation. [18], [19], [20]

Despite the definitions above, the term OD is not always used under this definition and sometimes it is broader. The five star deployment scheme is a format that classifies the quality of OD according to the format is made available in and considering the pros and cons of each "star level". Level one of the five star OD deployment scheme is simply described as "make your stuff available on the Web (whatever format) under an open license" [2] These two different definitions serve the purpose of raising awareness to the interpretation of the term OD since it might not always be applied with the same standards.

2.2.1 Availability, Status and Evaluation

The current definition of OD is that it should be open by default. [21] Openness and transparency from the government are not recent concepts. The right to access information is contemplated in article 19 of the United Nations Universal Declaration of Human Rights [22]. The WWWF takes it a step further by also defending that people should have the right to data. [16] It is not surprising that the inventor of the WWW, Tim Berners-Lee, was defending this back in 2012 by stating "Opening up data is fundamentally about more efficient use of resources and improving service delivery for citizens. The effects of that are far-reaching: innovation, transparency, accountability, better governance and economic growth." [23] and so was the vice-president of the European Commission in 2013 "Already today, governments - and the public they serve - are learning that open public data can boost transparency, improve public services and fuel

innovation." [24] The European Commission has shown to be committed to data-based initiatives, and it shows through the EU's investment in OD projects between 2007 and 2012 adding up to €454M, according to CORDIS. [25]

There is a consensus that OGD is not a question of if or whether but a question of how, by which means, with what standards, and what challenges. It is also asserted that making data available is the government's responsibility, alongside incentives to its re-use. [21] The latter is further discussed in the 2.2.3 Open Data Re-use chapter.

The Open Data Barometer (ODB), in 2017, studied the global state of OD. It set off with the principle that making data available to meet quotas is not enough. The first thing to look out for is what data is available. It was found that the public is more interested in data concerning the budget, spending, contracting, landowners, and company registers. Despite this interest, from all the datasets analysed by [16], only 22% fit these categories and they are often the least complete.

In different words but similar meanings, the OECD points out being user-driven, and proactive as digital government indicators and creating value through information and communication technologies as a goal. [21]

Table 5 Scores on Open Data Barometer 2017

Country	Rank	Score	Readiness	Implementation	Impact
Highest EU Score - France	3	85	100	71	88
Lowest EU Score – Croatia	58	27	52	24	8
Portugal	34	42	58	47	16
EU Average[1]1	-	51	65	49	38

Table 5 Scores on Open Data Barometer 2017presents an excerpt from the ODB which scores a total of 115 countries by the following parameters: 1) How ready the country is for OD initiatives (Readiness); 2) The implementations of OD programs (Implementation); 3) The impact it is having on society, business, and politics (Impact). Although, there is a connection between OD and corruption the top and bottom scores from Table 4 are not the same as these. Once again, there is no one size fits all, and the indicators are very different.

One thing that stands out is how low the score for Impact Table 5 usually is when compared to the other two. Denmark and Germany stand out for having balanced scores across the board. The average deviation of their scores is less than 2, they are also the only two countries with a value inferior to 4.9. This does

1 For the EU average Cyprus, Lithuania, Luxembourg, Romania, Slovenia, and Malta were not considered since they were not in the index.

not put them on the top, but it creates the appearance that both countries have controlled strategies to incentivize OD re-use. Portugal does not follow this example and if the countries were ordered by Impact the rank would drop to 49.

The status for OD can be summed as a work in (good) progress. The importance of it is understood, availability is the minimum of requirements and citizens are getting involved. The OECD includes OD in the second generation of citizens digital rights and the protection of personal data in the first generation. [15] OD walks a grey area of being useful for increasing public trust in the government while also needing people to trust the government with their data. This leads to data protection and security being the main concerns and challenges. [21]

From the 1 725 datasets evaluated by the ODB, only slightly over half were machine-readable, a quarter open-licenced, a third published with supporting metadata or documentation, and less than a third can be easily downloaded in bulk. [16]

Metadata is understood as data about data, or about the dataset, common examples are the source, date published, author, editor, scope, theme, and other similar attributes.

OD policies often come through as government platforms making datasets and documents available to the public. The OECD [21] created an eight-step model aiming to take policies from the bare minimum to the desired quality Attachment 1. When comparing the indicators the WWWF used in its barometer [16] with the ones from the OECD it is possible to observe that both agree on what makes good OD. The basic requirements are that it is free, findable, open-licensed, and proactively released. From here on, the investment should go towards encouraging re-use with the availability of bulk download, metadata, and API keys. It should also be time-sensitive, in the sense that it is up to date, and machine-readable.

OD portals need to be meant to create value, co-creation, and collaboration, and even if this is the case they are still not ideal. The ideal recommendation from both OECD and the WWWF is that governments should have one centralized portal. The goal would be for an interlinked central portal to be fed in real-time by data automatically uploaded from its source department [16], [21]. The WWWF specifies that the use of LOD and interoperable datasets would be a final stage where OD is no longer a goal but a means for a platform-based government.

In Germany “Article 12 in the law requires federal German authorities to provide data in machine-readable formats, with metadata descriptions, free of charge and with unrestricted access to re-use.” It also states that data should be open-licensed and proactively delivered. [21]

A good example of this evolution is the Transparency Register, an initiative from the European Commission aiming to give transparency to lobbying. The registry reveals what are their interests, who is

behind them, and what are their resources. This allows people to easily be informed about the activities developing and potential actors of influence. The register currently holds 12 000 entities, the growth happened mainly between 2013 and 2017, it is hypothesized that most relevant organizations have been included and any future growth is residual.[26]

Every organization, independent worker, and company whose activity influences in any way the politics of EU institutions must be registered. The registry and keeping it updated is what gives them access to the decision-makers, the EU Parliament, and the EU Commission. There are six sections of activity, 53% of the registries belong to in-house lobbyists and trade/business professional associations, 26% belong to non-governmental organisations, and the remaining four sections represent less than 10% each. This data can be exported in XML or Excel via the EU Open Data Portal. [26]

The EU Commission organizes several workshops and courses to spread knowledge about the database and encourage re-use. Quality control is also taken seriously, in 2019 the general secretariat analysed 4559 records. From this analysis, 46% were considered unsatisfactory and the entities were contacted to update and complete the information. This led to half updating the information and getting a green light and the remaining being removed. [26] This is a good example because it checks at least three big boxes: it has content quality control, an incentive to keep data updated and sanctions for those who do not, and activities to encourage re-use.

2.2.2 Portugal's Open Data and Transparency Scenario

Making every public act public with an open license and in a machine-readable format is a recommendation from TI to the Portuguese government since 2012. It was also recommended that every complex and long act would be accompanied by a summary. [22] The current state of OD in Portugal is halfway on these recommendations and that is what is going to be discussed in this chapter.

The Council of Ministers Resolution (RCM) n.º 108/2017. [27] approved an ICT Strategy for 2020 with the primary focus of making sure that data protection laws are being applied and digital security acknowledged. This responsibility falls on the National Commission for Data Protection. [15]

One of the ways in which this materializes is that citizens and businesses are entitled, to some extent, "to consent and refuse permission for the citizen or business data they provide to a given public sector organisation to be shared with and reused by other public sector organisations." [15]

These are important steps when it comes to compliance with GDPR (General Data Protection Regulation), but they are not as relevant when talking about OD and Data re-use. Portuguese legislation respects

international standards regarding openness and justification, but there still is a lack of implementation and citizen use of information. [22]

An example of the above is the portal (dados.gov.pt). It is inspired by the French version which is one of the "most highly developed central OGD portals" [21] and one of the reasons France ranks first in the EU and third worldwide in the ODB. [16] Datasets can be added by the public, these are identified to be easily distinguished from the ones published by public sector entities. Both are collaborative spaces which according to the OECD is a positive sign, even though it still raises quality control concerns from countries not so keen on this method.

The reality is that portals like this are helping Portugal score well in availability and accessibility, the big challenge remains re-use. On this topic, there are currently six data portals reusing data from data.gov. The biggest and worthy of mention are the health, justice, and municipal transparency data portals. [15], [28]

Criticism of the data portals has led to improvement before, as was the case of the Official Gazette, known nationally as *Diário da República*. The Electronic Official Gazette (DRE) initially published only pdf copies of the documents and had serious search limitations appointed. [22]. Currently, DRE has clear language summaries of complex legislation, the full text displayed on the website as well as downloadable in pdf. The legislation approving universal and free access to the DRE was only passed in 2016 by Decree-Law n.º 83/2016. [29] Although it is an example of criticism leading to improvement, it is important to note that DRE is a case of transparency not truly OD.

The public procurement portal Base.gov.pt has also been criticized for allowing the publishing entity to treat information before publication. [22] Public contracts are made available; however, the information is inserted by the entities, and it is not time-sensitive. There are publication dates years apart from the award dates and incomplete information on several contracts. It is hard to access what specific measures are being taken to address the limitations of Base.gov, but there are some leads. Law nº 3/2020 includes the plan to increase implementation of report obligations from public entities and allow the treatment of information concerning the excess use of directly awarded contracts already published in public portals (namely Base.gov). The same law also concerns modernization and better cross-referencing of interests declarations from high office politicians and public workers. [30]

Nevertheless, this is more focused on the content than on making Base.gov.pt more compliant with the definition of truly OD.

Portugal's OD situation is only positive regarding availability and accessibility, the third and equally important pillar of Re-Use remains neglected. This has not been made a priority in Portugal. In the OECD

score for government re-use of data support, in 2019, Portugal scored 0,06 at data promotion initiatives and partnerships, 0,01 in data literacy programmes in government and 0.08 in monitoring impact. The country improved between 2017 and 2019 both in availability (0.05) and accessibility (0.07), but not in re-use. In the latter, Portugal has the fourth worst score, tied with Finland. Re-use is Portugal's OD Achilles' Heel. Countries that managed to increase their score in re-use did so by promoting an OD culture and raising awareness on the benefits of reusing OGD. [21] From a technical standpoint, the availability of data and whether it is truly OD, in the sense that it is available for bulk download, machine-readable, time-sensitive and trustworthy, is an essential feature to incentive re-use. On the public side, in Portugal, there is a lack of evidence, sustaining whether citizens are ready or habilitated to act on the information available and create value with it. [22]

2.2.3 Open Data Re-use

The re-use and impact of OD is an important metric for both OECD [21] and WWWF [16] but is also the one with the lowest scores. In 2019, the EU updated a directive concerning OD and aligned with the General Data Protection Regulation, aiming to "encourage the Member States to facilitate the re-use of public sector data with minimal or no legal, technical, and financial restraint. In addition, the directive will make available high-value data for re-use." [31] Furthermore, opening up public data resources for re-use was considered strategic by the European Commission in their strategy until 2020. [25]

In more practical terms, data quality in the sense of non-biased, standardized, machine-readable, time-sensitive, proactive, and understandable data, is a prerequisite for re-use. [21] However, as was hinted in 2.2.2, it is not enough. Research shows that re-use needs to be encouraged by the provider, in this case, the government, in a proactive way. On one path, the public needs tools to take benefit from data, they need data literacy, awareness and programs illustrating the benefits of re-use projects. On a second path, these tools can be materialized in initiatives like hackathons, workshops, courses and cooperation and citizen-driven platforms. Value can only be created through the reuse for decision making, innovation, and providing information.

The prototype that was developed is essentially a data reuse project. Therefore, other re-use projects aimed at corruption-fighting are of interest to enrich the state of the art. The remainder of this chapter focuses on the comparison and analysis of such projects.

The starting point for this analysis is the European Public Sector Information Platform Topic Report No. 2014 / 04, entitled "Open Data as a Tool to Fight Corruption" which contains an "overview of how release and re-use of OD can help curb on a range of corruption forms in a number of sectors."[1]

For each sector of corruption at least one project is given as an example. These served as a starting point to find keywords and references that lead to other similar projects. This chapter explores the connection between the good OD practices laid out in 2.2 Background Analysis – Open Data and the projects that materialize some or all these practices.

First and foremost, let us settle on common ground. Although OD can be used as a corruption prevention tool it is not always aligned with this interest. [18] The databases made available to the public are not always the ones the public is more interested in, or the ones needed to create value. [32] The motive for making data available is also relevant. Often, OD serves as a way of showing transparency over what has already been done. This is where the difference between transparency and OD lays. Transparency can be achieved with OD, but it is only one of its many possible uses. For instance, for OD to be proactive and enable decision making it should be released and available for scrutiny before decisions are made, not after to show transparency over what is already done.

This is one of the reasons why it is important to create OGD that sustains and aids in decision making, having a direct impact on governance instead of being a final display of carried out actions. [33] An example of data that has the potential to aid in decision making is lobby registers, however, this is a rare dataset, for instance, a report looking into five G20 countries found that only one published lobby registers. [17] This is different from the EU lobby register mentioned in 2.2.1 Availability, Status and Evaluation which is not mandatory on a national level.

The impact of these projects and the amount created are hard to assess, in part because documentation on impact is lacking. [32] Most data re-use projects are government independent initiatives Table 6. An attempt to evaluate public interest in re-using data would also be faulty for two factors: 1) As seen above and again on these reports, training stakeholders to produce and use data is crucial. [32] As long as this gap exists, the access to data of public interest cannot be assessed; 2) As long as governments are creating only human-readable and digested data displays with attractive visualizations, instead of providing access to raw data, remains a barrier to re-use, and might not even pass the criteria for accessibility [21]. Creating final reports with charts is a very common practice useful to give the public an overall idea and a sense of transparency but it is not useful for exploring the data. Moreover, by the standards of the ODC, data visualizations are not OD, since they do not have the technical features needed to be re-used by anyone.

The remainder of the chapter dives more into the relation between OD re-use and the prevention of corruption, how to incentive and apply it, what different reports agree on, and the main challenges.

2.2.4 Open Data Standards

Good quality OD is emerging as an anti-corruption tool. On top of the requirements for OD established above, uniformization, standardization, machine-readable formats, and metadata are important to compare different datasets and look for patterns, trends, and anomalies that can expose corruptive practices. [18]

A comparison between the CPI and the ODB for 2016 (most recent ODB) shows a connection between the values in Figure 1 in the form of a strong correlation (0,71) between the two indexes[2]. A report by TI points out that OD "should contribute to anti-corruption reform" [17] and the ODC states that "Open data can play an effective role in dismantling corruption networks" [34].

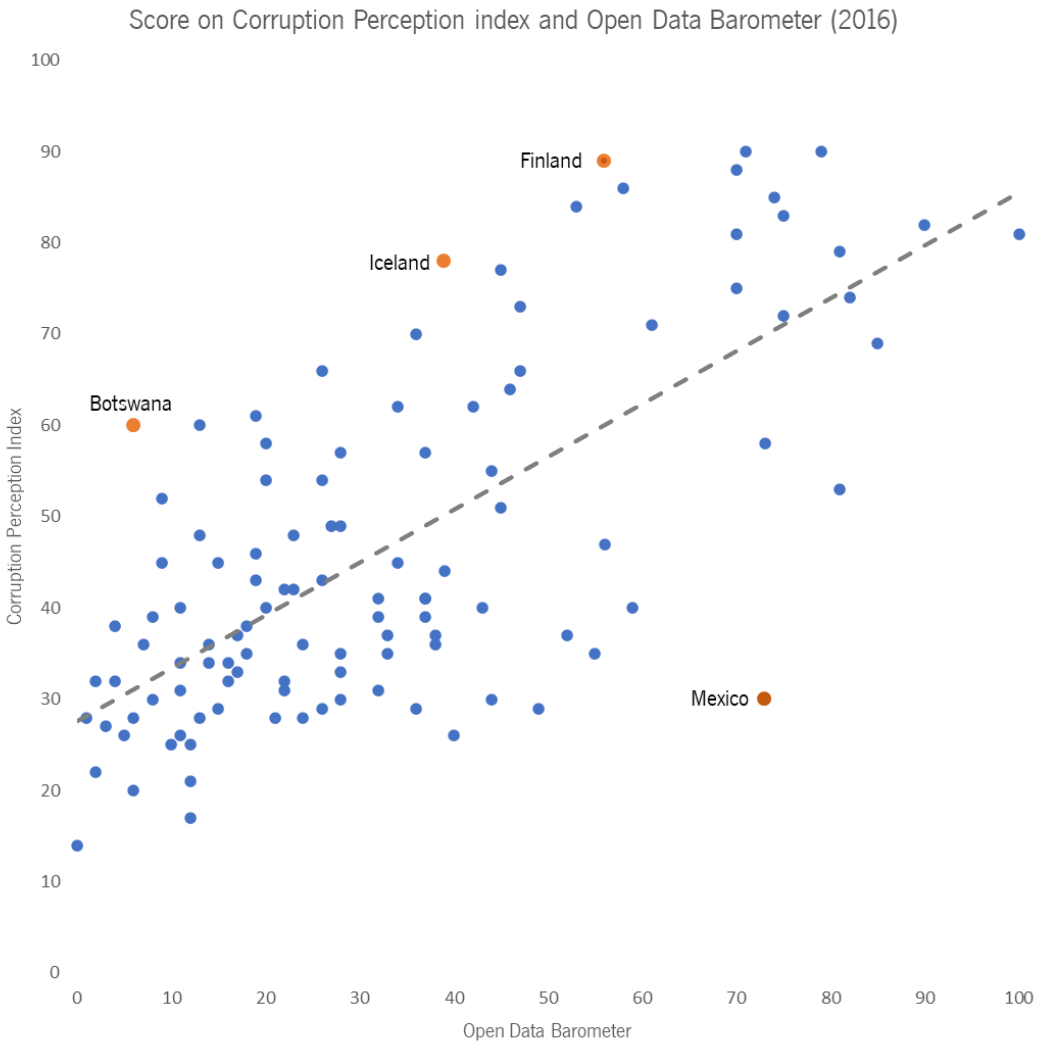


Figure 1 Scatter plot of CPI and ODB scores [17]

2 For this comparison only the countries scored in both indexes are evaluated, that makes a total of 113 countries.

OD is a vehicle for achieving transparency and citizen trust. These are values that impact the perception of corruption, however, it is not common for the impact of OD and the Perception of Corruption to be officially connected [33] as is showcased in Figure 1.

To further understand the relationship between OD and the perception of corruption, six different reports on the matter, ranging across America, Asia, and Europe[3], were investigated. The remainder of this chapter is based on these reports.

There is no recipe for corruption fighting, but there are good practices, and they are highly dependent on context.

The role of OD is tricky here. Theory points to transparency and OD being enablers of accountability, which is a way to fight and prevent corruption. But OD has not been widely used in corruption investigations. [33]

Despite it not being common, it is still a recommendation. The G20 established a set of principles directly connecting the anti-corruption measures and OD [18], and the ODC makes it clear that OD has an important part to play in the anti-corruption reform. [17]

The impact of OD in anti-corruption is related to the third pillar from the OECD, data reuse. This is the stage that allows for further investigation from third parties which are then able to mine data for incorrect practices. [33] These are even more powerful when the publishing format allows and incentivize triangulation, linkage, and comparison. [17]

Besides the technical features of the data, the content is also important for re-use. In these scenarios, more data is not the same as better data.

Certain datasets are more relevant for anti-corruption than others. The “Open Up Guide” divides these priority datasets into three types: registers, public disclosure, and transactions. Amongst the total of 30 datasets, the top 5 are all registers, namely: interest declarations, lobbying register, company register, charity register, and politically exposed people's list. [34] The G20 also listed 10 key datasets [18] that are of the public interest [34]. Below in Table 6 are the top ten datasets from the Open Up guide and the ten datasets from the G20.

3 [32] Malaysia [1] EU [20] Latvia, Sweden and Finland [18] Brazil [19] Indonesia [35] India

Table 6 Priority datasets [34] and key datasets [18]

Open Data Charter [34]	G20 [18]
Interest declarations	Lobbying Registers
Lobbying register	Company Registers
Company registers	Beneficial Ownership Registers
Charity register	Directories of Public Officials
Politically exposed people's list	Government Budget Records
Public officials register	Government Spending Records
List of government contractors	Public Procurement Records
Corruption-sensitive postings	Political Financing Records
Council / advisory board members	Legislative Voting Records
Contracts register	Land Registers

The four shaded titles on each column represent the datasets that are both in the list from the G20 and the top ten from the ODC.

From the six key datasets in the right column of Table 6 that are not shaded, only the Beneficial Ownership Registers is not mentioned at all by the ODC.

These recommendations are extremely valuable. Opening government data is a process, and, since it cannot be done all at once, institutions must prioritize release. With this guide, prioritization can be done strategically and governments should be choosing the datasets known for creating greater impact upon release. [1]

One thing is knowing what, in theory, should be available to apply scrutiny on the public sector. Another thing is knowing if it is working. The reports analysed all showed similar approaches and challenges. Amongst the main challenges identified the one that pops out the most is lack of training. This applies both to public officials and the general population. [20]

When evaluating datasets, the challenges were also quite similar. Overall, the published datasets are not always the ones needed and the formats are not easy to re-use resulting in a lot of time spent in data treatment and transformation. [32]

The conclusions are also similar, in Indonesia, a report states that there is an effort, but it is not enough. The country needs a dynamic platform to create value. [19] An analysis in Brazil adds that functioning legislative frameworks on Data Protection and the Right to Information are key for policies to evolve and transparency to become cultural. [20]

The main challenges related to re-use are the technical and content quality of the data and the data literacy in civil society. The first challenge reflects itself in that low quality and non-centralized data makes it harder and more time consuming to follow connections, names and stories across different sectors [33]. As for data literacy, it is connected to the technological gap affecting public officials, investigative

journalists and civil society that disables them from using data to prevent and detect corruption. [17], [18]

2.2.5 Data Literacy

The lack of data literacy has been repeatedly mentioned in the last sub-chapters as a challenge. Although there is also no quick fix for this technological gap, the paragraphs below display some examples of both training sessions and incentives to data re-use that have been successful in different countries.

Finland, Latvia and Sweden formed a Coalition to foster the use of ICT technologies that resulted in both training projects and programs to raise awareness in civil society. [20]

In Brazil, several e-learning courses, online resources for government staff, and guidelines on how to use OD are only some examples of the tools available to boost the impact of OD.[18]

In Malaysia, the "Sinar Project", on a limited budget, managed to create consumer applications and research notes based on their combined database. The database connects politicians to the companies they own, and the constructing deals these companies landed. The project includes data on 4000 people of interest, 200 organizations/committees, 900 government posts and 222 current parliament members. This project was expanded to allow exploration of local politicians, private and public sectors. [32]

In India, "Project Background" displays arranged data visualizations for the whole country and, one instance on a state level (Maharashtra). This project used data from different national and international databases to create a diagnostic tool. The description was completed with the outcomes of a Mumbai workshop to include public opinion. The remarkable thing about this project is that it includes public opinion as the main part of it and allows for replication on any state inside or outside India as long as they hold a workshop similar to the one in Mumbai. [35] The model proved to be scalable, as promised, through successful implementation in Mexico and Mexico City.

In conclusion, the main challenges related to re-use are the technical and content quality of the data and the data literacy in civil society. The first challenge reflects itself in that low quality and non-centralized data makes it harder and more time consuming to follow connections, names and stories across different sectors [33]. As for data literacy, it is connected to the technological gap affecting public officials, investigative journalists and civil society that disables them from using data to prevent and detect corruption. [17], [18]

2.2.6 Open Data Re-Use Projects

All the OD re-use projects mentioned in the introduction to 2.2.3 Open Data Re-use are analysed below in Table 7. The analysis is part of this project and aims to better understand how the scenario for data reuse laid out above materializes in online projects.

Description of Table 7:

Title – This is the name of each project. Corruption area – One or more areas are attributed to each project. The areas were picked according to [1] which based them on the ODB and describes them as "forms of corruption and affected sectors as dictated by leading anti-corruption indices and surveys. (...) it rather focuses on sectors and corruption-related issues that are most pressing and evident."[1] Then, there are six Boolean variables where (X) means it exists and (blank) is not found. Due to language barriers in certain projects, the information had to be translated online. This means that there is a possibility that some (blanks) exist but were simply not found. Each of these six fields is now further explained:

Is Active – Whether or not the website is up, and the information updated. "Diferentonas" is the only case on which the website is unavailable, but the information is found in the report. The others have the website up, but the information is old or discontinued.

Private Initiative – Boolean on whether the project is a private initiative or public. All public initiatives are governmental on either a local or national level. Public initiatives are still data re-use projects when they feed on other portals. For instance, "Portal Municipal" feeds on Dados.gov.pt.

Data source – Whether the data sources for the projects are specified so it is possible to trace the data re-use back to its source it is also considered true if it is made evident that the data is generated by the project. OD Re-use may be partial.

Source Code – Whether the source code is made available. Almost every true case has a repository on GitHub[4].

Shows Method – Whether or not the website explains how the data is treated and analysed. Some positives are more specific than others almost allowing replication.

Has documentation – Whether articles (academic or not) and reports were found about this project. It is very helpful since reports tend to be very comprehensive and justify the means and the gap the project came to fill.

Data Display Mode – Most of these are projects that handle a lot of data and considering the principles of OD mentioned above it is important to evaluate how the user data is displayed. This column shows

4 https://github.com/

how the data can be accessed and if it enables further re-use. For instance, visualization projects may draw an interesting conclusion but create a barrier for further work with the same data.

Table 7 Analysing projects

Title	Corruption Area	Active	Private Initiative	Data Source	Source Code	Shows Method	Has Docs	Data display mode
Influence Explorer	Political parties Business Lobbying		X	X		X		API & Bulk
Abgeordneten Watch	Parliament / Legislature	X	X				X	Visualization & Q&A
STIRNA	Media Corruption	X	X	X				Visualization, JSON, XLXS & CSV
Open public contracts in Slovakia	Business Public Procurement	X	X	X			X	API (paid) & Bulk
Open Courts	Judiciary	X	X	X	X	X		API
Rekvizitai	Public Officials Civil Servants Public Procurement	X	X	X				XLXS and CRM integration (if requested)
Tender	Public Procurement		X	X		X	X	Visualization
Project Background	Public Procurement	X	X	X		X	X	Visualization
As diferentonas	Public Procurement		X				X	-
RECORD: Reducing Corruption Risks with Data project	Public Procurement	X	X	X				Visualization (links source)
Portal Municipal	Public Procurement	X		X		X		CSV
OpenSpending	Public Procurement	X	X	X	X	X	X	API, CSV, bulk, Visualization (personalized)
Where does my money go?	Parliament/Legislature	X	X	X	X	X		Bulk
Vouliwatch	Parliament/Legislature	X	X				X	CSV, Visualization & Q&A
Lisboa Aberta	Public Procurement	X		X				CSV
Votaciones – La Nacion	Parliament/Legislature	X	X	X				CSV & Visualization
Dollars for docs – How industry dollars reach your doctors	Public Procurement	X	X	X		X		Visualization & CSV (paid)
NarcoData – Animal Politico	Public Procurement	X	X	X		X		Visualization & Tabular
AtuServicio.uy	Public Procurement	X		X	X	X	X	Visualization (compare)
Open Contracting	Public Procurement	X	X	X	X	X	X	Tabular, JSON, CSV
Quien es Quien	Business, Lobbying	X	X	X	X	X		API & Visualization (filter)

The sample of projects selected turned out to support the conclusions depicted above.

The most prominent corruption is depicted in Public Procurement, which is in focus on 12 of the projects, it is followed by Parliament/Legislature appearing in 4 projects.

When it comes to the Boolean variables, being a private initiative (18) and displaying their sources (18) are the two most common characteristics. Amongst the first, there are mainly newspapers, third party corruption fighting agencies, and independent citizens. The source code of the project is, unfortunately,

only available for 6 projects. Having some kind of methodology description is a more common practice taken by 12 of the projects. Nine projects also disclose documentation including both reports and news articles.

From the 21 projects most showed only visualizations and only four had an API. These projects are, at least partially, re-using OD, which means that is available on its source. Nevertheless, it is a good practice to make your data available, so others can replicate similar projects or triangulate the project's data with new datasets.

All in all, this short analysis allowed for two things: understanding the practical side of the OD challenges and recommendations discussed above; designing the first sketch of the final project based on what exists online and what is working with people.

A very interesting project is "Abgeordneten Watch" which was founded in Germany and lead to the creation of the Parliament Watch Network5, from which "Vouliwach" above is also a part. This project consists of a website with information on members of the German Parliament (data re-use) and a question and answers feature where the public can directly ask questions to the politicians, and they can respond. The implementation was very successful, and the project is active. Further down in the "Methodologies" chapter, the impact of this project on defining the structure for the project developed in this thesis is further explained.

OpenSpending is a free, open, and global platform to search, visualise and analyse fiscal data in the public sphere. It comprises over 3,527 data packages from 85 countries with over 187,226,528 fiscal records.

NarcoData is a tool that recompiles trusted information on the last forty years of drug cartel history in Mexico. It is managed by Animal Politico, a native digital Mexican newspaper. It is especially interesting because it touches on a topic that runs in the shadows and back streets and about which it is not easy to find organized and structured data.

"Quien es Quien" is about contracts between public entities, businesses, and politicians. This tool makes it easier for investigators, journalists, and the common citizen to access OD from different sources hence enhancing transparency and accountability. This is a pure re-use project that has the simple aim of making research easier.

The four projects singled out in the previous paragraphs are not necessarily the best of more complete projects, they are simply examples of variated the application of OD can be.

5 https://parliament.watch/

2.3 Linked Open Data

"Info that is not linkable is not used, information that is not used is not valuable" [36] This is a bold statement and one that is fit to start the discussion about LOD.

2.3.1 Semantic Web

First things first, the concept of LOD is preceded by that of the semantic web. The latter was introduced by the creator of the world wide web, Tim Berners-Lee, and refers to the existence of a web that is readable both to humans and machines. What this means is that the web as we know it today is limited, it provides access to enormous amounts of information, but this information has already been processed and isolated. To compare it to more online available information a traceback must be done and do all that work is manual. This is the online equivalent of roaming a library for different books mentioning the same theme. The break of barriers and the unleashed potential of the internet are in being able to have all this information linked and available for comparison. This is what is discussed in this chapter.

Having semantic in the web context essentially means that information is represented in a graph-based data model that facilitates extension, integration, and inference. [4] Giving semantics to data is the process of linking it to more data to be as understandable for a computer as a written text would be for a human. The preferred data format to achieve this goal is RDF, which allows the links connecting data to be described. Ultimately, the connection between documents is implicit, something that is often lost in regular data formats found on the Web of Documents. [37] The web of documents is the web of today, with separate websites and pages showcasing their own information. The Web of Data is an extension to the web of documents applying semantic principles. So, they naturally share some properties: There are no restrictions on to what data types can be contemplated; Everyone can publish data; Publishers can choose their vocabularies; Entities are connected by RDF links. [37]

From a development perspective, in "Linked Data – The Story so Far", the web of data is further described: Data is strictly separated from visualization aspects; Data is self-describing; HTTP is a data access mechanism and RDF a data model; Applications are not stuck to a fixed set of data sources, new data sources can be discovered at run-time by following RDF links. [37]

Understanding the Web of Data is understanding the ultimate, long-term goal of LOD implementations. LOD is a set of best practices, and those are what is discussed further on.

2.3.2 Linked Data

Linked Data refers to "a set of best practices for publishing and connecting structured data on the Web." [37] In practice, it means that all data is linked to instances or connections from different sources. It can be applied on an organizational level connecting databases that are not interoperable, yet. Or, on a trans-organizational level linking data from different organizations in a unique shared database, as is the case of data sharing in Alzheimer's disease research. [38] The Alzheimer data share case started from the following premise: Professionals agreed that sharing data on Alzheimer research was valuable but did not know how to do it because every organization's internal database had a different structure. They found in LOD a method for tagging data from different organizations around the globe with the same tags and lists of terms, therefore, having access to more data and supporting better research.

This short description shows that, by nature, the requirements for LOD present a solution for some of the barriers identified above by open-data re-use projects. Namely, the challenges related to scattered non-centralized data.

Back in 2006, Tim Berners-Lee defined a set of four base principles for publishing LOD on the web: Use URIs as names for things; Use HTTP URIs so that people can look them up; Provide useful information on URI's using the standards (RDF, SPARQL); Include links to other URIs. [37] Overall, they lead to one of the main cornerstones in LOD, interoperability. In Table 8 the main technologies needed for understanding LOD are explained. These technologies are "semantic", so the information is "represented on a graph-based data model that facilitates extension, integration, inference and uniform querying." [4]

Table 8 Main Linked Open Data Concepts Explained based on [37] and [39]

URIs	Uniform Resource Identifiers use the http:// scheme to identify any entity in the world. These entities can be looked up by dereferencing the URI over the HTTP protocol. [37]
RDF	Resource Description Framework "A graph-based data model that facilitates extension, integration, inference and uniform querying." [4] RDF is based on triples: subject (S), object (O) and predicate (P). S and P can be strings but should be both URI's and O is what connects them. For instance, S is the mother of P. "is the mother" is the object that gives context to the link between the two. [37]
SPARQL	*SPARQL* is the language used to query an RDF dataset and the semantic Web. A SPARQL endpoint is a service that processes SPARQL queries and returns results. [4] Its logic is similar to SQL. It enables any user to query the web as if it is their local database. [40]
Vocabularies	"Vocabularies provide the semantic glue enabling Data to become meaningful Data."[41] They are essential to LOD. Whenever possible well-established vocabularies such as FOAF, SIOC, SKOS, DOAP, vCard, Dublin Core, OAI-ORE or GoodRelations should be used to make the

	connection between data easier to understand. “Vocabularies are themselves expressed in RDF, using terms from RDFS and OWL, which provide varying degrees of expressivity in modelling domains of interest.” [37] Vocabularies can either express properties and classes or describe values.
RDFS	RDF Schema is a language for creating vocabularies. This step should only be taken if no vocabulary already in use can describe the connection needed. This is the last resource measure. If deemed necessary the new vocabularies need to be connected to existing ones and completed with RDFs or OWL definitions. [37]
OWL	Web Ontology Language is another language for creating vocabularies. “It is designed to represent rich and complex knowledge about things, groups of things, and relations between things.” [42] OWL-based technologies allow for incremental model growth and permit freedom of changes and adjustment throughout the development phase.. [27] For instance [43] uses isPartOf and its owl;inverseOf hasSubsequent to offer better browsing experience.
LOV	Linked Open Vocabularies is a dataset of vocabularies that allows for free text searching for vocabularies. [4] It does not have every existing vocabulary and does not include value vocabularies, only properties.

These concepts are used multiple times throughout the following chapters and in the solution development. The use of the resources presented in Table 8 should be aligned with the best practices that must be considered in any LOD project. Some of them are: Publishing the metadata of the dataset in RDF and using common namespaces wherever possible; Using vocabularies to describe data without ambiguities and facilitate reusability, interoperability, and data quality assurance;

Using common terms (such as agency names, government sectors, and parts of countries) to automatically generate common URIs. [44], [45]

The last point raises the importance of agreeing-upon data vocabularies in the scope of each theme to facilitate data interoperability.[45]

Designing an LOD application, in the sense of the properties and value vocabularies that are used, and how the objects are connected is an essential step. The Dublin Core Metadata Initiative (DCMI) is leading an effort in this subject, the Metadata Application Profile. The DCTAP is an ongoing project led by the Dublin Core Community whose main goal is to develop a “Simple Tabular Model for Application Profiles (AP-STM)”. [46] This translates to the standardization of the profile development process profiles, in a simple and human-readable tabular format that can then be used to generate machine-readable validation schemas. [46] The project was born to satisfy two confined community interests: developing tools to aid in creating and documenting application profiles; Having application profiles that specify validation rules for the data they define.

A profile is essentially the description of the data in terms of defining the types of subjects (rdfs:Class) their expected properties (rdf:Property) and expected values (rdfs:Datatype, skos:Concept). DCMI's proposal is in a tabular format which should be filled bearing in mind the project's guidelines [46] and the RDF and RDFS rules. LOV as presented in Table 8 can be of great help in finding the properties and classes but should not be the only search engine used.

The alignment of the DCTAP TAP with validation rules raises an important subject. Before RDF data is made available to the public there should be a validation step. In other words, a way to ensure that the data is conformant with what is described in the Application Profile. The DCTAP TAP is built considering this future step using Shape Expressions (ShEX) as the validation language.

ShEx is a language for describing the shape of RDF triples. In other words, a schema is composed of shapes each shape being the equivalent to a subject type with its intended properties and intended values of said properties. ShEx enables the definition of subjects, predicates, objects, cardinality, and datatypes. The level of specificity of a shape schema is dependent on the author.

A ShEx schema may be used for validation purposes where RDF data is parsed against its intended shape and is either conformant with the shape, or non-conformant. [47]In a simple example, if a shape defines that every subject of the type foaf:Person must have exactly one foaf:name and at least one foaf:mbox address, then any subject of the type foaf:Person with, no foaf:name, more than one foaf:name or no foaf:mbox is not conformant.

The intended use of Schemas includes validation, communicating the data structure, transforming RDF graphs, and generating user interfaces. [47] In the scope of the solution, only the validation and the communication uses are approached.

Validation is ShEx is structure-wise, it does not include the quality of the data in terms of content.

After validating the structure of the data, it is important to offer different formats for data exploration to include people in every stage of the data literacy knowledge curve. Ideally, the platform must allow link exploration and SPARQL queries. Also, implementing faceted-browsing capabilities, metadata search features, and structured-query Web services is a good practice. So is the idea of making other data formats available by appending to the URI. URI.json returns a Java-Script Object Notation and URI.csv returns a comma-separated values description and so on. [44] Another powerful suggestion is the use of an API. [27]

2.3.3 Advantages of LOD

It is widely agreed that the main advantage of LOD and LD is that it offers raw data and the possibilities that derive from it. LOD is easily customized to the needs of every user. [40] Making available LOD formats rather than "read-only" displays increases collaborative participation since more data can be added and linked to include different sources, discover patterns, and customize applications. [44] The natural structure of the triples allows for the creation of a unified database, diminishing data dispersion and isolation. RDF is also, by nature, suitable for multilingual data through language tags that easily allow for data to be described in multiple languages simultaneously. [48]

Just like the language tags, metadata, data about data, is also integrated. Instead of being a separate file or description from the dataset, LOD is self-descriptive. "Linked metadata are just additional triples that are stored together with other data triples. First, it allows publishing metadata on data level and second, it enables querying metadata and data at the same time." [3]

It is never too much to remind that being machine-readable does not mean it is only machine-readable. LOD is not restrictive in that matter. As seen in the best practices, it is possible (and recommended) to allow bulk dataset download and to have links to and from other LOD or LD projects.[44] Other visualization techniques are also encouraged.

From a data management perspective, when compared to Entity Relational Model (ER) many differences pop up. When exploring an ER database, one needs knowledge about how the data is organised. In opposition, LOD formats display data in a way that resembles the natural language (subject, object, and predicate is a simple sentence structure). This model is universal, and the interpretation of data is aided by ontologies and semantic context. This leaves little room for local specific denominations (unless absolutely necessary) or interpretation struggles. [40]

One of the main criticisms of non-linked data is the creation of what is known as "data silos". This concept applies to OD sources that co-exist as silos scattered across the internet with different formats, structures, and semantics. They hold valuable information but said information only communicates with data from the same silo. [49] Web APIs are great for the re-use of the database they are applied to (the equivalent of a silo) but do not assign globally unique identifiers to data items. This makes it impossible to link items in different data sources. "In contrast, Linked Data applications can work on top of an unbounded, global data space (...) connecting the different data silos that currently exist on the Web back into the single global information space." [37] This is essentially the possibility to use the web as your database.

2.3.4 Linked Open Data Implementations – With a Focus on Government Data

Research done by Avila-Garzon in 2020 on 250 LOD papers found that the main areas of implementation have been biology, social sciences, libraries, research, and education. Government data is considered among 14 other topics in social sciences which take the second largest piece of the total research (20%), closely followed by biology (19.2%). The largest slice is Other (34.8%) and includes 27 sub-topics. [50]
Moreover, the LOD cloud[6] has grown from 12 datasets in 2007 to 1,255 datasets with 16,174 links in 2020. [51]
Government data is selected here as a focus area because many of the datasets that are of interest to the public Table 6 belong to, and therefore must be made available by, the governments.
The Rensselaer Polytechnic Institute has a team focused on transforming government data into an RDF format. Their work includes developing tools and techniques for interacting with government data and linking government information to other resources. They have hosted 60 online demos using technologies for visualization, APIs, and Web service composition. They also offer tutorials on how to mirror these demos. [44]
The ontology inherent to an LOD application is ideal for the representation of hierarchical and complex relations, as is the case in government-related issues such as e-procurement. [27], [48] The results of the research conducted over the impact of the Netherlands' Cadastre Land Registry and Mapping Agency, which is (2020) the largest implementation of Linked Data in the governments of the Netherlands [3], concluded that Linked Data can support the business vision of governmental organizations and should be part of business discussions, including vision and ambitions statements.
On a larger European scope, LinkedEP is an LOD project applied to the European Parliament analysed through a report. [43] In the 29 weeks following the announcement of the data output the team received over 5000 visits and 7504 queries, being 3850 of which made with SPARQL. An inspection on logs also shared that the most common SPARQL queries involved regular expressions (43%), count functions (42%) or both (24%). To increase usability, the backbone of the model is a direct translation of the structure of the events in Parliament [43]
The above mentioned is one of the examples of EU's work towards a semantic web, but there are many. The EU is a great source for controlled vocabularies in the form of Thesauri, Authority Tables, Taxonomies, and others described in SKOS about several themes. These are published on the official website for the Publications Office of the EU [52]. One of the most complete vocabularies is EuroVoc. It is a Thesaurus covering information about the activities of the EU described using semantic web technologies. These

[6] https://www.lod-cloud.net/

vocabularies are a great asset because they come from a reliable source, have the tendency to be widely implemented and are usually available in many languages, EuroVoc, per instance, is available in 23 languages. [52]

Although from a financial perspective, releasing data in its raw format "as is" is cheaper than rendering it into reports and applications, just having the raw data is not enough since it implies the following work of cleaning and standardizing data before being able to perform any analysis. [44]

The potential for an LOD model in government data is that the raw data is available regardless of other visualization techniques applies. Therefore, transparency becomes a consequence of a well-oiled machine, not the goal. This potential is best represented in the following example inspired by a similar phrasing found in [49].

For the question "How much did companies owned by politicians make from public procurement?" One would start by querying a search engine aiming to find the answer, maybe in an article. Since that is unlikely, the procedure to find out would involve going through asset declarations, one by one, to find out what companies are owned by politicians. Then go through the public procurement portal looking for contracts for all these companies. Finally, the data would have to be analysed to sum the amount of money earned and answer the question. Along the lines, other challenges could appear, such as having different names for the same company. However, if these two datasets, both from governmental sources, were published with LOD practices, this whole process could be easily replaced by a SPARQL query.

Unlike the examples in 2.2.6 Open Data Re-Use Projects which were mainly re-uses of already online data by third parties, the LOD projects found are more commonly official implementations. Some of those LOD projects in the range of OGD are described below. Reviewing these projects is important for the building of the solution, especially when it comes to choosing vocabularies.

GovWILD: Integrating OGD for transparency is a prototype that integrates and cleans OGD on a large scale. The goal is to offer a database that answers questions about politicians, companies, and government funding. The project re-uses data from ten different sources, out of which only three are not HTML. The motivation is simply to showcase the power of this OD once connected. [49]

Data.gov is America's largest government LOD database offering mostly five-star OD, and in some cases, four-star. [2] At first, it was planned as a data publishing and access platform. After noticing that the public was more interested in specific areas instead of general data, they changed the model. Now, data.gov is the main structure automatically feeding portals known as communities where people interact with the data they most care about. Main communities are about health, semantic web, business, oceans,

law, energy, education, public safety, and research and development. Their triples use terms from linked data vocabularies and provide links to the data sources. [44]

LinkedEP is an LOD implementation that translates government data into RDF. It took inspiration from LOD parliament projects existing in the member countries that originated in three groups: Governmental projects; Civic parties; Academia. They aim to link to these projects and ultimately become what they call a "Web of Linked Government Datasets". [43]

The CLAV platform is a contribution to the availability of public administration OD in Portugal. It offers services oriented to public administration, citizens, and companies. The team behind CLAV highlights an ontology linking businesses of entities with public office, the legislation regulating these processes, and the ones responsible for executing and preserving information generated along with these processes. The project also offers detailed metadata for all integrated documents and aims to widen its audience by displaying their data in an open format at Portuguese OD portals and integrating with the European catalogue. Currently, CLAV allows data to be downloaded in CSV, RDF, OWL, SKOS and XML. It also supports exploration with SPARQL, integration with LOD or Data.gov. and the use of an API key. Everything in this project was done following the General Data Protection Regulation (GNPR). [27]

2.3.5 LOD Challenges

"The task of providing unified, structured, and interlinked data is daunting but worthwhile. Published clean data can be analysed, visualized, or further interconnected. Amongst others, a benefit is heightened transparency of government actions." [49]

The first challenge, especially when talking about data re-use, is finding good quality organized data. The quality varies largely, from different formats for bulk download to websites that need to be crawled to retrieve data. On top of this, the schematics of different sources are also different and inconsistent. This poses a great barrier for integration, turning normalizing and treating data to a global schema into a complex and time-consuming task. [49]

The first challenge is the lack of a standard methodology for implementing LOD. The papers evaluated by [50] lead them to recommend further investigation to focus on developing a validated standard methodology for managing LOD. Most studies implemented their own steps, and most failed to report clear phases of their methodology. Combining all studies, eight methodological steps were identified: Interlinking; Annotation; Publication; Retrieval; Content Generation; Transformation; Storage; Visualization. No study used them all and only one used seven. The runners-up completed only three and the most implemented step is interlinking. [50]

The second challenge is the LOD knowledge curve challenge. There are plenty of tools out there but most of them are not user-friendly for querying LOD datasets without knowing a query language such as SPARQL. [50] SPARQL knowledge is considered a limitation since it is not a widely dominated skill, its potential for next to unlimited queries ends up being limited by the need to learn the language. [43] Moreover, the lack of general knowledge about LOD itself ends up also being a barrier for the users to take advantage of the full potential of this data model. [50]

However, usability can be prioritized for non-specialized users through the implementation of natural text search and other possibilities. Another technique, used by [43], was creating redundant properties to enable less complex and shorter queries and avoid reasoning engines. This practice is ill-advised for large datasets.

The third challenge is asserting trust in data. Therefore, it is so important to connect to known vocabularies and publish metadata. Linking helps to assure maintenance and reliability through being transparent about data provenance. [37]

To conclude, four main challenges need special attention: Data quality and availability; Lack of a standardized methodology for implementation; LOD knowledge curb; Data Trust.

2.3.6 Implementation

Although there is no set methodology, there are guidelines commonly followed. "How to Use Linked Data" [4] proposes a checklist for creating LD. This checklist, represented in Table 9 is a decomposition of the three basic steps for implementing LD suggested by Bizer, Heath, and Berners-Lee. [37] the three basic steps are in the first row separating the columns.

Table 9 Linked Data Checklist from "How to use Linked Data".

Creating LD	**Interlinking LD**	**Publishing LD**
Verify that all relevant entities/concepts were effectively extracted from the raw data?	Link dataset to other RDF datasets	Provide dataset's metadata. Provide information about licensing.
URIs are dereferenceable	Created terms are linked to other vocabularies	Create alternative access methods: SPARQL endpoint; Data dump.
Terms used are from widely accepted vocabularies. Only non-existing terms were created.		Register the dataset in LD catalogues via a Data Hub or CKAN.

This checklist helps guide the implementation process. Despite not being mentioned above, one last step that can improve this checklist is the addition of other Visualization techniques. [50]

When doing a re-use project, the step that comes after selecting the data sources the first step is extracting the data from said sources. The complexity of the next step depends on the format of the data file (XML, CSV, JSON, SQL, PDF). For sources with data dumps, a common procedure is simply standardizing all the sources to the same data format file, for instance, JSON. For other web sources with no data dumps a common practice is crawling the websites for the information needed. This can be aided by tools such as GATE, Zemanta and DBpedia Spotlight. The result of this procedure is then transformed into the same type as the remainder data. [49], [3], [4]

Depending on the formats available the next step may be cleaning and normalizing the data or jumping forward to transforming it to an RDF format. The latter is a crucial part of any LD or LOD implementation. This is when vocabularies are chosen to enrich the data, relationships between triples are defined and properties are chosen. [48] Increasing interoperability relies largely on the choice of well-known and used vocabularies like FOAF and Dublin Core. [43] The relations defined should also have inheritance between them, this can be partially achieved with the use of properties such as rdfs:subPropertyOf and rdfs:subClassOf to relate properties or classes from different vocabularies. On the schema level, RDFS, OWL and the SKOS are useful for the same goal. [40], [4]

Another important thing is, when possible, linking to external knowledge sources already in the LOD cloud. [43] One of the main tools to do so is DBpedia. [50] It is important to remember: LOD can never be isolated. This process of transformation is tricky but there are tools that aid in the transformation of both tabular data (OpenRefine[7]) and ER (R2Rml[8]) to linked data. [4]

Moreover, whenever possible the URIs should be human-readable, using specific references to entities. When working with LOD the application profile should not be restricted by local usage samples. The data should be published “as is” with as few logical constraints as possible to allow flexibility.

For instance, when defining an Application Profile it would appear acceptable stating that “the triplet: "author is Academic Member" requires that any author is personal of the institution.” [40] This would be acceptable in local implementations, and restraints such as this are common in ER models but the reality is that, in many cases that restraint does not apply. [40] In LOD restraints that have a local or project, the specific scope should be avoided to increase re-use of the ontologies.

7 https://openrefine.org/
8 https://www.w3.org/TR/r2rml/

Next is the implementation phase. Here we must remember the importance of metadata to describe the dataset, element definitions, value lists, provenance information, construction and transformation methodology, and potential use cases, and limitations. [3]

After the data source is translated into RDF, it needs to be made available for further consumption. When everything is ready to be published, a specialized LOD server such as Virtuoso is recommended. [50] This step enables querying the data with SPARQL. It is also recommended to implement enhanced services based on semantic methods, visualization techniques and other tools to increase human usability and improve the final user's experience. [3], [48]

The process is complex and there are numerous guides and recommendations for implementation. For that reason, this chapter did not go further into exploring tools. That is done later in the methodologies with everything being applied to this project's case scenario.

3 DATA AND METHODS

Inside Data And Methods lays the description of every decision made designing the proposed solution and how it was based on the state of the art. From the data sources to the tools and methodologies. In Figure 2 Concepts Map there is a recap on the concepts presented in the State of the Art, how they are connected and the navigation throughout them that led the researchers to the tools and materials.

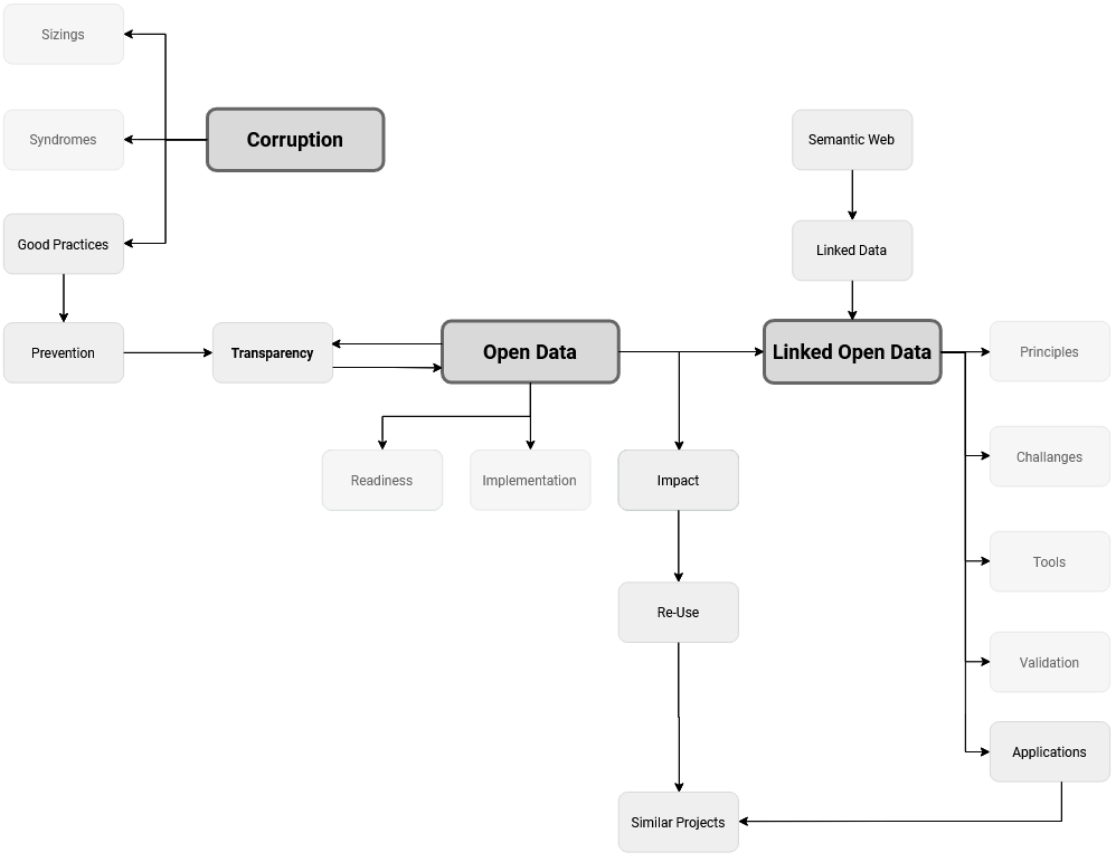


Figure 2 Concepts Map

The state of the art starts with the definition of corruption, how it is measured, the syndromes, and the good practices to fight it. For the scope of this project, the focus is on good practices, in specific prevention. Literature on prevention showed that OD is a corruption-fighting tool through scrutiny enabling transparency, hence the title of this document. The State of the art also showed that one of the measures for the impact of OD is the level of Re-Use and that the highest level of OD Is LOD.

A considerable number of projects defining themselves as an "Open Data Project to Prevent Corruption" were analysed. As for LOD projects making the same statement, not many were found. Nevertheless, in spite of not having the prevention of corruption as a stated goal, the topics of the governmental LOD

projects analysed are in line with the areas of interest for Open Data in the prevention of corruption defined by the ODC.

Before moving to the characterisation of the proposed solution it is important to point out that this is not a one-way path. Meaning that prevention is not the only way to tackle corruption, transparency is not the only way to act preventively, and OD is not the only way to achieve transparency. It is, however, a path well documented and sustained by literature and the one that best fits the scope of this work. For context and research credibility the state-of-the-art chapters put these concepts against their neighbours and alternative paths.

This information is summarized in Figure 2 in a concept map resulting from the research done above. The proposed solution places itself in a gap found in the literature: There are few Open Data Re-Use projects with machine-readable and linked data.

The data selected is aligned with the datasets of interest found in the state of the art for corruption and transparency, which means it is a scrutiny enabling project. The technical field is the semantics web, more specifically LOD.

3.1 Method - Design Science Research

This project follows Hevner's three-cycle approach to design science research. [6] The author states that "practical utility alone does not define good design science research. It is the synergy between relevance and rigour and the contributions along both the relevance cycle and the rigour cycle that define good design science research." [6]

From a broad perspective, the idea behind the original model Attachment 2 A Three Cycle View of Design Science Research [6] is that there are three pillars: Internal Environment; Build Design Artifacts and Processes; Foundations. They all iterate with the pillar immediately before and after through the Relevance and Rigor Cycles. The Build and Design pillar also iterates with itself through the Design Cycle. In 2016 the optional External Environment pillar and Change/Impact Cycle were added to the model [53] as an expansion to the original three cycles of 2007. [6] This expansion is not used because interaction with the external environment is out of scope for this project.

The guidelines behind this approach date back to 2004, [54] and each is thoroughly explained therefore aiding in applying the model to any project. Table 10 Design Science Guidelines according to Hevner shows every guideline applied to the project at hand.

Table 10 Design Science Guidelines according to Hevner

Guidelines	Map4Scrutiny Domain
Guideline 1: Design as an Artifact	The main artifact is the final linked open data dataset.
Guideline 2: Problem Relevance	Stakeholders' concerns about corruption aligned with readiness and potential of Open Data.
Guideline 3: Design Evaluation	Linked Data Principles, Open Data Quality, Shape Validation, and Querying capability.
Guideline 4: Research Contributions	Improve the impact and scrutiny potential of already Open Data by transforming it to LOD.
Guideline 5: Research Rigor	Methods used in other Linked Data applications with adjacent themes, best practices, and guides.
Guideline 6: Design as a Search Process	Multiple-step and prior-knowledge based implementation.
Guideline 7: Communication of Research	The proposed solution is open to the public and articles will be published.

Each guideline is an important step of academic work. Design as an artifact created the final product, whether it is a theory, a framework, or a model. In this case, it represents the final LOD implementation. Every other guideline either justifies, adds value, or reflects the impact of the artifact. The initial idea of using LOD as a tool to fight corruption was broad and had little supporting research. It took a high level of abstraction, going back to the isolated concepts of LOD and Corruption to understand if there is a connection, if this connection has causality and how it can be approached.

Below, Figure 3 shows Hevner's model applied to this project with the information from the guidelines above in Table 10. The Background and Gap Analysis on both Corruption and OD established the need for projects of this nature along with an existing connection with LOD, the extent of which is further explored in the Internal Environment. As showcased in Figure 2 the research goes from an overall search including three different concepts to a specific prototype.

A key factor in fighting corruption is preventing it, which can be achieved through transparency that can be achieved with OD, which's the highest quality is LOD. Topic-wise, Portuguese citizens, one of the

stakeholders, are interested in data about the activity of politicians and public officeholders. Thus, the proposed solution will transform OD that fits these criteria into LOD. This implementation is also aligned with the notion of F.A.I.R Data. Data that is Findable, Accessible, Interoperable, and Reusable.

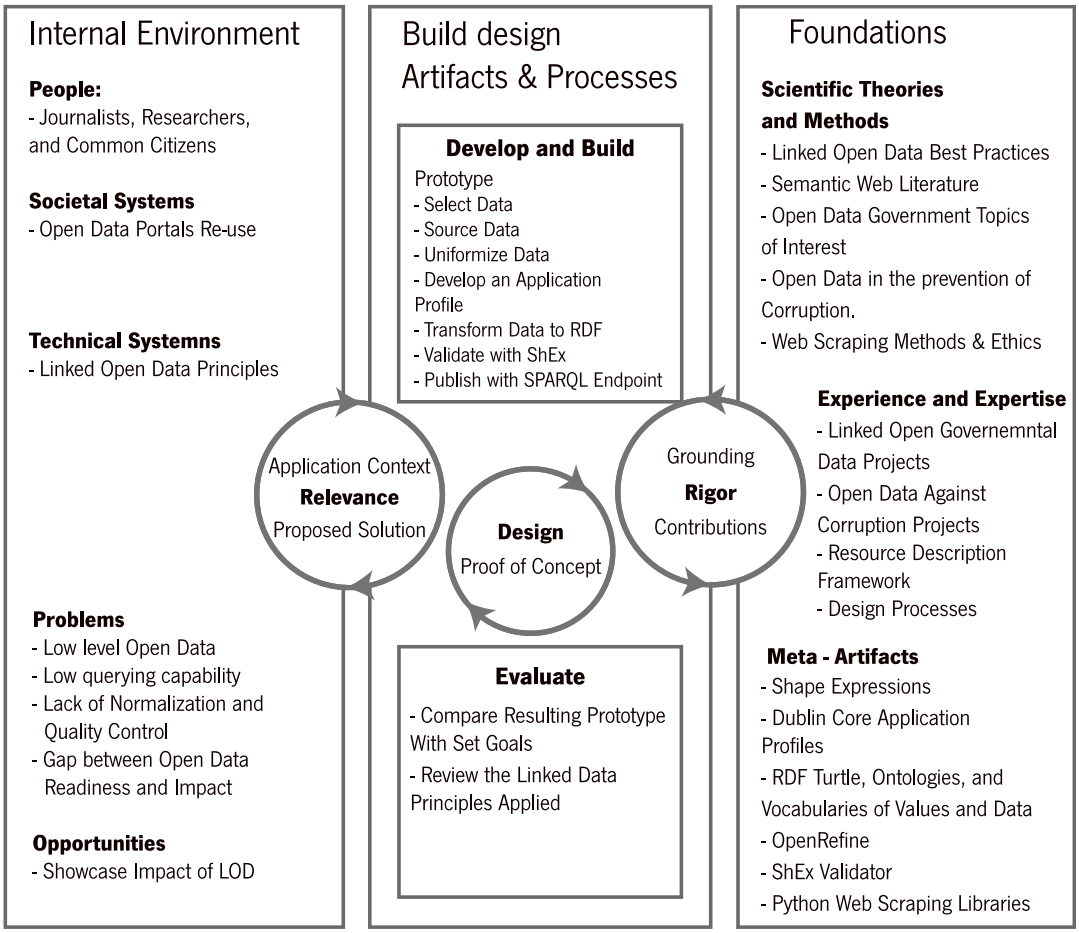


Figure 3: Drechsler, Hevner (Ed.) 2016 - A three-cycle view of design science research [6]

Going from left to right in Figure 3 we have the three cycles applied to this project. The people, societal systems and technical systems set the idea for the project. In this case, we are talking about the first half of the background research that shows how corruption is a concern if it is of interest to the public, and that OD is a possible solution, but most implementations lack quality or incentive to re-use.

This takes us to the problems and opportunities where the scope is defined. Ideally, this project aids in sustaining the arguments for publishing OD with better quality and in machine-readable formats. Thus, showcasing the potential relevance of the proposed solution. These are the requirements for the chosen datasets, the main problems, and the opportunities. [6] Understanding this environment is crucial to designing a proper solution in the next pillar. Here the topics of interest to the public and the barriers to the existence of similar projects are the focus points.

The building pillar is where the design and build of the proposed solution happen. This pillar alone iterating itself multiple times is the application development. It is its connection with the other two that justify the relevance and create knowledge. The work being developed must remain relevant to the internal environment and improvements must be explained and measured. This is the difference between design science research and simple application development. [53] All the work happens within this cycle, but every decision is based on knowledge brought from the rigour cycle.

The relationship works both ways. Existing projects contribute with methods and knowledge and this project contributions back to the knowledge base for future projects. The foundations include inspiration sources, for tackling opportunities/problems, existing artifacts, analogies/metaphors, and theories.

"Research contributions to the knowledge base are key to selling the research to the academic audience just as useful contributions to the environment are the key selling points to the practitioner audience." [6]

By following this approach, one can create and develop while keeping a good grasp on research, inspiration, and relevance. The pillars iterate with each other multiple times in each step of the development. A sample of the result of this iteration is the use of web scraping and reification techniques. These are mentioned but not further explored in the state of the art because the need to use them was caused by the characteristics of the selected data. The use of these tools was not planned before research, it was a consequence of the first iterations between the second and the third pillars. The need was felt by an obstacle in the design cycle, the solution for the obstacle was found in the foundations and therefore included.

The evaluation on good practices and technical correctness of the approach is done by basing the decision making in literature: Is the dataset considered "Of Interest" by the ODC?; Is the Linked Data checklist complete? These two questions are answered based on Table 6 and Table 9. The conformance of the data transformed with the application profile is validated via ShEx by parsing the data against a defined shape schema.

At the bottom of the last column, the tools gathered are part of the knowledge base to build the solution. These contribute back to the knowledge base in the form of output files that help future researchers in following the same approach. These output files also materialize the objectives set in the Introduction.

3.2 Proposed Solution - Description

The Overall design and implementation of the solution fit in the second column of Hevner's four-cycle view of design science research portrayed in Figure 3.

3.2.1 Data Sources

Before the LOD implementation, the first step was defining a data scope for the prototype. This is subdivided into two more specific steps: Narrow the choice down to a few databases; Select what data is being used from each selected dataset. Considering the importance of data provenance, trust, the public interest, and the concept of F.A.I.R data that is explained above, the following list of requirements for the datasets emerged:

- Open Data is free to access, extract and Re-use.
- About entities or activities susceptible to corruption.
- Of interest to the public.
- About Portuguese entities.
- Aligned with the key datasets presented in Table 6

Table 11 Seven Up to Date Open Databases in Portugal

Name	Focus	Search Options	Export Options
Base.gov.pt	Public Procurement	Keywords Search bar and filters	Bulk with multiple data and PDFs for contracts
dados.gov.pt	National Open Data Portal (multiple themes)	Keywords Search Bar and Navigation through categories	Bulk and API
Pordata.pt	Municipalities, Portugal, and Europe (multiple themes)	Navigation Through categories and Advanced Filtering	Bulk (data can be filtered before exporting)
transparencia.sns.gov	Health	Keywords Search Bar and Navigation through categories	Bulk datasets
estatisticas.justica.gov	Judicial and Law	Navigation through categories	Mostly PDF's
dadosabertos.turismodeportugal	Tourism	Navigation Through categories and filters	Bulk and API
INESCtec	Research (multiple themes)	Keywords Search bar and filters	Bulk
INE	Social Statistics (Multiple Themes)	Keywords, Navigation Through categories, and Filters	Bulk and API

Since no document was found with a list of the main or most relevant databases in Portugal, the databases portrayed in Table 11 are the ones often mentioned in European documents (dados.gov.pt, and Base.gov), laws, articles, and the ones that are easy to find by simply querying a standard search engine for OD Portugal.

From the bottom up, INE is the national institute for statistics and offers a diverse collection of data on the most varied social and demographic topics. Most datasets are available for bulk download and the source is credible. Data from this portal is essential for any project that needs data on the demographic properties of Portuguese society. INESCtec is not of interest since it is not governmental or directly related to entities or activities susceptible to corruption. The Tourism portal feeds on Dados.gov.pt, therefore, it is irrelevant for this project. Justice.gov has data on judicial decisions, court cases, and more. The fact that most data is available only in pdf is a disadvantage since it would mean taking data all the way from level 1 to 5. Transparencia.sns.gov is very organized, has a large variety of datasets about health, and makes most of them available for bulk download. Pordata.pt is a private initiative lead by the Francisco Manuel dos Santos Foundation. It has very powerful embedded visualization tools and allows the cross-reference of different tables from their dataset. They also offer metainformation with every dataset and allow for bulk or a filtered download. The other two: dados.gov.pt and Base.gov, are the most mentioned Portuguese data portals, they have been mentioned in the Open Data Chapter as good examples, which does not rid them of limitations. Dados.gov.pt has better usability and an API, on the other hand, Base.gov.pt has data on public procurement and fits all requirements set at the beginning of this chapter. Base.gov.pt is chosen as one of the main databases for this project, which is now lacking a second database and the connection between them.

Some municipalities' open data portals were also found by the same keywords search, OD Portugal. They are not included in Table 11 Seven Up to Date Open Databases in Portugal because they are not centralized in the sense that each municipality publishes their data for public scrutiny separately.

With the key datasets from Table 6 at hand, the next step was to search for these datasets in the order of priority to understand if they are made available in the country. The idea is to link to Base.gov.pt above a dataset as high as possible in the ODC hierarchy. The first searched term was "Interests Registers" which returned Parlamento.pt. This is a website focused on the activity of the Portuguese parliament with a page dedicated to OD with XML dumps. As for the Interest Declarations, those are only made available in HTML or PDF depending on whether they belong to members of parliament or ministers.

Exploring the portals mentioned above and looking for other sources of OD led to the draft of two possible prototypes that could have added value if improved and linked to Base.gov. The two drafts are broadly described in Table 12.

Table 12 Prototype Scope Options

Base.gov.pt + Parlamento.pt			
Focus	Main Connections	Main Source	Data Availability
Parliamentarians	Parties, Spouses, Public offices, and connection to organizations.	Parlamento.pt	Data is available without the need for authentication.
Organization	Public contracting, social value, and headquarters.	Base.gov	Bulk download per organization.
Pros		Cons	
This is of public interest to the point that similar projects already exist in seven other countries connected by the Parliament Watch Network.		Data on parliamentarians is available as HTML, it must be extracted and treated.	
Possibly interesting queries: Parliamentarians with both public and private offices, by political party. Public contracts of organizations connected to parliamentarians. Spouses inside Parliament. Accumulation of positions per party or region.			

Base.gov.pt + Municipalities			
Focus	Main Connections	Main Source	Data Availability
Municipalities	The investment plan, public contracts, use of EU money, spending plans.	Pordata's data on municipalities and their publications on individual websites.	Available online without the need for authentication but the quality of the data is very dependent on each municipality.
Organizations	Public contracts with municipalities, social value, and headquarters.	Base.gov	Bulk download per organization.

Pros	Cons
Public interest There are municipalities with advanced data publication platforms such as Lisbon. There are guidelines on how to connect and map data under this theme.	A prototype could not use all municipalities (like in Project Background [35], and this hurts the potential for comparison. Quality of the data in some municipalities.
Possibly interesting queries: Geographical contrast between the municipality and contracted organizations. Compare the cost for similar contracts in different municipalities. Compare the distribution of funds per municipality.	

From the standpoint of public interest and the existence of similar projects to base the approach, both ideas are doable. Therefore, the decision came down to which idea is more appropriate for the timeframe of a prototype implemented in a dissertation.

Sourcing the data: All interest declarations are presented in the same way, even though they are HTML they can easily be scraped. On the other hand, not all data on municipalities is on Pordata or Dados.gov.pt and the data from individual municipalities would have to be retrieved from each website individually and every website is different. This means that it would take longer to source data on municipalities than parliamentarians.

Size of the dataset: There are 230 parliamentarians versus 308 municipalities.

These two are the main reasons why the link with the parliamentarians' interest registers was chosen over the link with municipalities. The remainder of this chapter dives further into detail of what are interest declarations and what kind of data is available on Parlamento.pt.

On a closing note, on this two-path topic, it is important to note that being the prototype of an LD implementation it is possible to expand into the other topic in the future while maintaining the links to the parliamentarians.

3.2.2 Base.gov.pt and Parlamento.pt

In Table 6, the first dataset mentioned by the ODC is the "interest declarations". A google search for Interest Declarations returned Parlamento.pt[9], a website that holds the interest declarations for all members of the Portuguese government (70), independent administrative entities (7), and parliamentarians (230). The declarations from the first two are available for download in PDF, for the latter they are displayed in structured HTML pages, one per each parliamentarian.

These declarations include the companies in which they have worked, still work, or are shareholders. On top of this dataset being classified as important, the success of the work developed by the Parliament Watch Network, mentioned above in 2.2.6 Open Data Re-Use Projects, served as inspiration and incentive to classify Parlamento.pt as a fit contender to link with information from Base.gov. The linking entities are the companies according to the following rationale: The parliamentarians are the first piece to be described. Each is/was a member of none or multiple organizations. Describing every parliamentarian returns a list of all the organizations linked to at least one parliamentarian. These organizations are then searched on Base.gov.pt which returns the public contracts the organization was involved in if there are any. If contracts are found, the contracts it is involved in are also described.

The resulting dataset shows the power of linking, querying, and making available data from two different sources that are both governmental. In Table 13 both sources are categorized in the same way reports by TI characterized open governmental data sources. [18], [20].

Table 13 Categorization of Datasets applying the same model as [18] , [20]

	Data	Time	Granularity	Format	Openness	Accessibility	Open Standards	Metadata	Documentation
Interests Declaration (Parlamento.pt)	X	X	X	X	X	X			
Public Procurement (Base.gov)	X		X	X	X	X			X

Starting with what they have in common, both sources are open and free to re-use, asking only for credits. The data is online and requires no login or other information before access. Everything is detailed and assessable. Parlamento.pt has an advantage in being timely and offering some datasets in CSV and XML but none of these datasets are the interest registers in discussion for this project. On the other hand, they

9 https://www.parlamento.pt/

also have data available only in PDF and HTML. The formats of the data are varied but the existence of some bulk downloadable files could be a positive indicator for further improvements in the future. Parlamento.pt has a sitemap to aid navigation through its multiple pages, but even so, navigation is not straight forward and the topics of the pages not clear.

Base.gov.pt is more user-friendly with a very straightforward navigation and user experience. The website was updated in March 2021 and now includes more detailed search query filtering, and the possibility to export a CSV with a maximum of 2000 lines of positive results. Each line is a contract. It also makes documentation available but lacks metadata. One of the main limitations is data not being timely. Details on contracts are uploaded by the contractor and uploads have been found about year-old contracts. Both databases were carefully examined in the sense of understanding what data could be taken from each of them.

3.2.3 Interests Declaration

Parlamento.pt has a page named "Open Data" which has datasets in XML and CSV. Despite being a legal obligation for the interest registers to be public, they are not part of those datasets available for bulk download. According to Article 26° in Law n. 60/2019, it is mandatory for parliamentarians to deliver a declaration of their profits, properties, and interests in the same terms that apply to all owners of high public or political office. Furthermore, the law states it is the duty of the Portuguese Parliament to ensure that the data on the interests is to be made public on their website. [55] The website states that the register is public and available for consulting in their portal. [56]

The fact that this dataset is not published as OD, creating an incentive to reuse, is a limitation for this prototype, but no legal grounds were found for this limitation to exist.

3.2.4 Selecting Data

Similar to what happens in other re-use projects, not all data available on the sources is used [43]. Selecting the data was a process of balancing between not simply using everything, and risking giving narrow perspectives due to the lack of data. The final dataset prototype is meant to allow people to explore and query the data, not to give them a biased vision.

For this reason, the data portrayed in the interest registers is enriched with the addition of some biographical markers. Following the example of [43], every parliamentarian has the official ID, the first and last name, birth date, electoral cycle (was country of representation in LinkedEP), affiliations to

committees, and political parties. The biographical data was downloaded in bulk (XML) from the OD page in Parlamento.pt and then appended to the scraped CSV. This data was structured and easy to append to the CSV registers by using the full name of the parliamentarians. This process is further explained in chapter 4 Implementation.

As for the attributes in the interests' registers, everything displayed in each register is an attribute because omitting information there creates the risk of having biased data. This is especially relevant for fields that leave space for comment.

As for Base.gov, the CSV export is very helpful and nearly every value was kept. In addition to the downloaded files, the links to the entity pages on the website were also included.

In Appendix A and Appendix B, there all the attributes from both sources are listed with some details on their content and a marker on whether they were kept or discarded, and a reason for the latter. These two documents represent the process that led to the final attributes used in the proposed solution and is represented in Table 14.

Table 14 Selected Attributes

Person (Parliamentarian)	Entity	Contract
• Full Name	• Name	• Contract Object
• Birthdate	• Nature and working area.	• Procedure Type
• Civil Status	• Type of role	• Contract Type
• Partner's Full name	• Headquarters	• CPV's
• Marital Regimen	• NIF	• Buying Entity
• Education Level		• Service supplier
• Job		• Agreed Price
• Parliamentary Commission		• Publication date
• Party		• Contract Celebration Date
• Parliamentary group		• Execution time (days)
• Electoral Cycle		• Place of execution
• Legislature		• Grounding
• Roles		• Extinction Cause
• Social Shares		• Contract Closing Date
• Had a Role In		• Actual Bulk Price
• Start Date		• Reasons for changes in timeframe
• End Date		• Reasons for changes in price
• Accumulation		• State
• Support or benefits		• Framework Agreement Number
• Services		• Framework Agreement Description
• Exclusivity declaration		• Centralized Procedure
• Other statements		• Link to Procedure Pieces
		• List of Suppliers

4 IMPLEMENTATION

As was mentioned above, there is no set methodology for implementing an LOD approach. Knowing this, the tools displayed below, and the implementation steps are based on the work of Avila-Garzon [50] that identified technologies used in several LOD management processes, and the Linked Data Checklist from "How To Use Linked Data" [4] that is available in Table 9.

Roaming back to Figure 3: Drechsler, Hevner (Ed.) 2016 - A three-cycle view of design science research [6] with Hevner's three-cycle view this chapter refers to the third column "Build Design Artifacts and Processes".

Overall, the phases of an LOD implementation are Creating, Interlinking and Publishing. To better explain what work is done on each phase this chapter makes an overview of the three main processes that are then described in further detail. The three main processes are in Figure 4: Data Sourcing, Cleaning, and Uniformization; Designing the Linked Data App Profile; Transformation, Validation and Publishing.
Phase one, creating, takes in the first and second processes, the phase two, interlinking, is designed on the second process, and implemented on the third, lastly, the publishing phase occurs entirely on the last step of the third process.

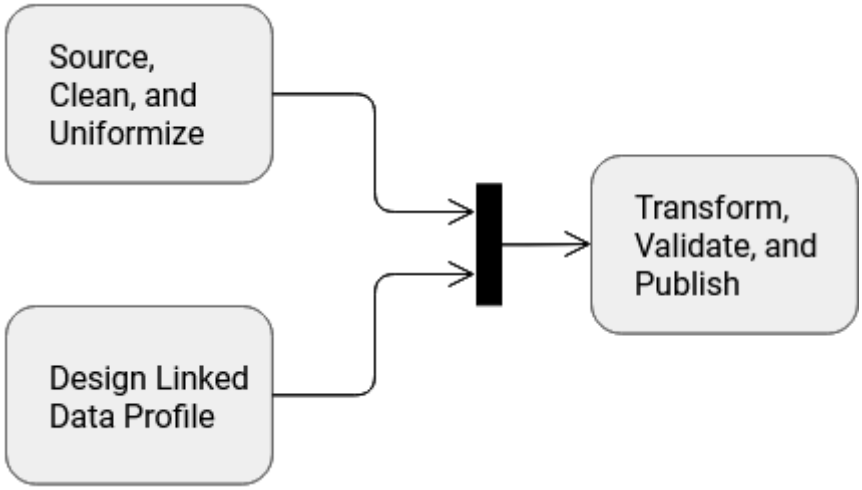


Figure 4 Data Journey - From Sourcing to LOD

Figure 4 represents a Macro perspective meant to illustrate the data journey from the source to being published as LOD. The important idea to take away from this view is that the steps in Source, Clean, and Uniformize and the steps in Design Linked Data Profile happened simultaneously but separately. Meaning that no outputs from any of the processes were needed to continue the other one. They did not interact with each other but both processes needed to be finished to start the Transform, Validate, and Publish process.
All decisions and steps are now described in the following sub-chapters: 4.1 Data Sourcing, Cleaning, and Uniformization, 4.2 Designing the Linked Data App Profile, 4.3 Transformation, Validation and Publishing.
To aid in understanding the processes there are a set of Activity Diagrams below following the norms presented in chapter 15 “Activities” of the Unified Modelling Language Specifications [57].

4.1 Data Sourcing, Cleaning, and Uniformization

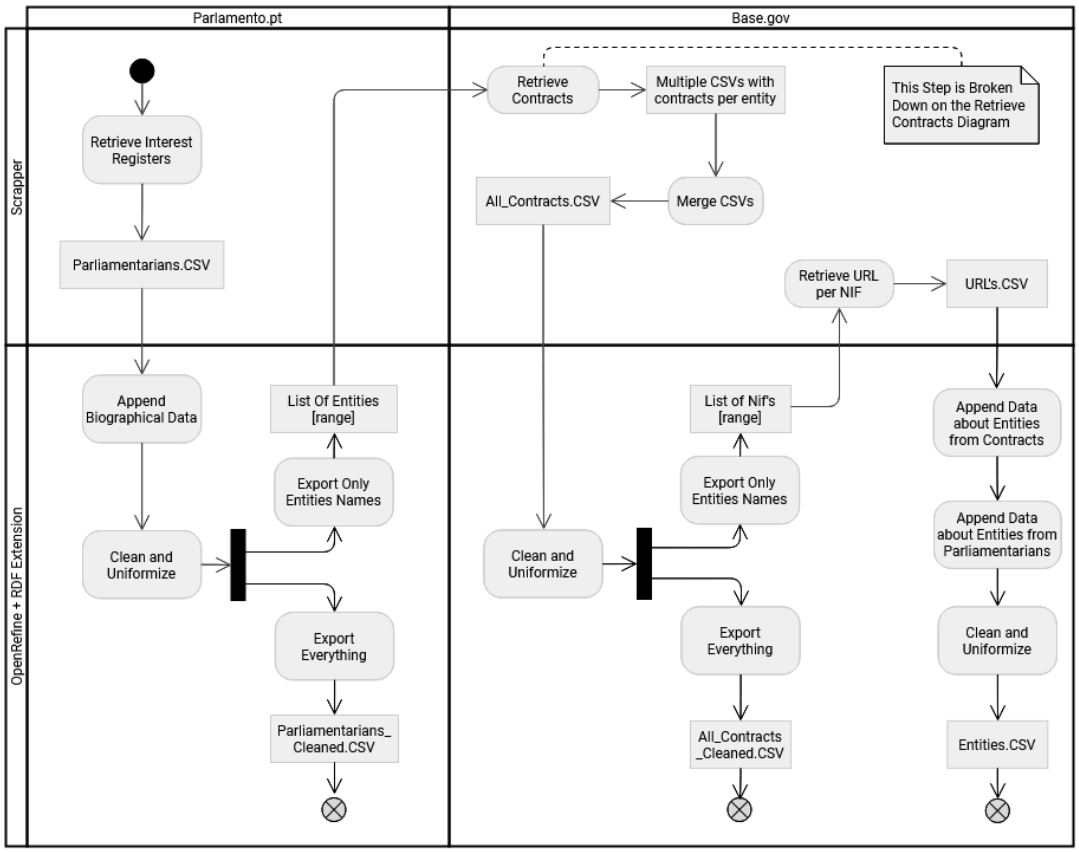


Figure 5 Source, Clean, and Uniformize

Ideally, a transformation to LOD would start from Tabular files or an ER database. However, since the data selected from Parlamento.pt is in HTML, the first step is to scrape/retrieve it and turn it into a CSV. Figure 5 shows the journey from the Data Sources on top "Parlamento.pt" and "Base.Gov" to the cleaned tabular data (CSV format) ready to transform into RDF. On the left side, the swim lanes divide the Scrapper, and OpenRefine + RDF Extension. The first refers to every Python script to retrieve data from the web and the second, OpenRefine, is a recommended powerful tool for transforming data, reconciling, and exporting in RDF formats with the extension of the same name. [4]. The course of the activities flows throughout the sources and the tools to better highlight which data came from where and how it was handled.

The steps here are both chronological and causal, one cannot be done before its ancestor. Below, 4.1.1 Sourcing explains the processes behind the activities in the scrapper's horizontal swim lane, and 4.1.2 Cleaning describes the processes behind the Open Refine + RDF extension swim lane.

4.1.1 Sourcing

The process of data sourcing describes everything in Figure 5 Source, Clean, and Uniformize that crosses the swim lane "Scrapper" with "Parlamento.pt" and "Base.gov". All the code written for this purpose can be found in the support files.[10] It is written in python with aid of the libraries listed in Table 15.

Automating the data extraction not only avoids human error but also helps others re-use the process and update the data in the future due to new elections, updated documents, changes of the statute, or switches in parliament seats.

Table 15 Python Libraries

Library	Purpose
Beautiful Soup	Parses the data from XML/HTML and navigates the parse tree.
Requests	Sends HTTP requests.
Selenium	Interacts with browsers like a puppet. Needs a web driver to run and is very useful with ASP.Net pages.
Pandas	Aids in reading/writing tabular Data frames.
Timer	Creates sleeping times to avoid overloading the servers with requests.

Web Scraping was born from the need to analyse data from the internet, and the incapability to do so manually due to the large volume of available information. It essentially is the practice of automating and systemizing the extraction and organization of relevant data from the web to analyse or combine said data. [58], [59]

Scraping data from HTML structures is also not new for LOD implementations. In [49] they used crawlers with built-in text extraction for sources without dumps, and OpenRefine itself offers a simple built-in scraper that parses the HTML of any given URL [60].

The practice is as old as the internet itself. Currently, setting up a scrapper is straightforward and requires minimal programming effort. [59] The most commonly used languages are R and Python largely due to their multiple libraries destined for this purpose. [58]

For most scrappers the macro process is defined in three phases: website analysis/navigation and access, website crawling/parsing relevant data, and organization/output [58], [59].

The first phase is analysing the website (or websites) to understand how the scrapper is going to be built, from here, a software agent mimics browsing interactions and retrieves the data needed. This is done by

[10] https://doi.org/10.34622/datarepositorium/KWFSU2

parsing each page. The last phase is retrieving and keeping only what is necessary and organizing the data, usually giving it an output. [58], [59]

Naturally, the process of automating the extraction and reusing data published by others raises discussions over the legality and ethics of web scraping. Throughout the years, practitioners have come up with a set of good practices and indicators to look out for that are always presented next to articles and tutorials on the matter. In [58] Krotov and Silva identify a preliminary set of ethical and legal questions and considerations based on both legal and information systems literature that aid in keeping the scrapper harmless to a website. Even in technical books, such as "Web Scraping with Python" there is a chapter dedicated to the good practices for ethical Scraping.

All the formal and informal articles, books and tutorials mostly agree on the main aspects that keep web scraping legal and ethical.

Respecting Robots.txt – These files cannot be enforced, they have no blocking power, scrapers can read and obey them or go around them. [61] That being said, the latter is likely not a good idea. This is a simple syntax text file that states what directories of the website the developers want or not to be accessed by robots. This file should always be on the first level of the website tree, this is also how it can be found, by appending "/robots.txt" to the URL. These files are very useful for scrapers since by telling you "where" not to go the developer is implying he is aware of the existence of web scraping and accepts the practice anywhere else on the website, except for the denied directories. [61]

In this project - Parlamento.pt has no Robots.txt file and Base.gov.pt has one as of the launch of the new version. This file states a 10-second delay for all crawlers and a few out of access URL'S. Both conditions are respected by the code developed.

Avoid Overloading with Requests – Just because code can go through all the pages of the website and get the needed data at a rate of a page every 2 seconds, it does not mean it should do it. It is always best to delay the scrapper and let it run longer, so it mimics human interaction and avoids overloading the server with requests. The consequences for not doing this could be damaging a website and that can be problematic for the author. [61] Despite the possible problems, overloading a server with requests is never morally correct regardless of the consequences. If the website is primarily accessed by people in the same time zone (case of both sources) it is also better to run the crawler late in the night when it is less likely to be used by humans.

On top of these two basic practices, Krotov and Silva [58] came up with a set of questions to which negative answers mean there is a "decreased likelihood of legal problems and ethical controversies in

their work". If there are negative answers it does not mean Scraping is completely out of the table, there might be an alternative solution such as contacting the owner of the website and getting permission. The questions and their answers in the scope of this project are presented in Table 15.

Table 16 Questions on legal and moral pointers for data Scraping [58]

Questions	Base.gov	Parlamento.pt	Justification
Legal			
Is Web Crawling or Web Scraping explicitly prohibited by the website's "terms of use" policy?	x	x	Neither have explicit Terms of Use that need to be accepted to navigate the website.
Is the website's data explicitly copyrighted?	x	x	There is no obvious copyright information.
Does the project involve illegal or fraudulent use of the data?	x	x	Anyone can access the data without registration or payment.
Can crawling and scraping potentially cause material damage to the website or the Web server hosting the website?	x	x	Scrappers are slowed down and run at night to avoid any damage.
Ethics			
Can the data have obtained from the website compromise individual privacy?	x	x	Does not apply because the data used is open and the public has the right to this information.
Can the data obtained from the website reveal confidential information about the operations of the organizations providing data or the company owning the website?	x	x	It can reveal connections but not confidential information or trade secrets.
Can the project requiring the Web data potentially diminish the value of the service provided by the website?	x	x	Financial losses and concurrence do not apply since there are no ads, and profit is not the goal of making this information linked.

All questions on legality and ethics in Table 16 Questions on legal and moral pointers for data Scraping [58]have a negative answer in both sources. None of the questions in legality applies to either of the sources. As for ethics, the data used is meant to be public, the project is only adding context by linking existing data making the datasets interoperable. As for the last question concerning the value of the site, the website is always the source and the project states that clearly by redirecting to the source. For this reason, the prototype is not considered to affect the value of the websites.

Some other important practices to apply when developing a scrapper are:

- When an API exists, it should always be prioritized over a scrapper. [59] For instance, Twitter's robots.txt prohibits access to directories containing data that can be obtained via their API. [61]
- Scraping information that is of open access to any human is unlikely to be a legal problem. The same does not apply to information protected by passwords or other means of access. [61]
- When it comes to the risk of copyright infringement this usually only applies to creative work, therefore leaving out company names, prices, names, and any other facts. [61] The data used in this prototype is already Open by Default, so the protection of the information does not apply.

With all of this in mind, two scrappers were written for this project. One that navigates through Parlamento.pt and retrieves data from the parliamentarians' interests registers, and another one that looks up companies on Base.gov.pt and returns the contracts they are involved with when there are any, then looks up the NIFs of all entities involved in the returned contracts and returns a link to their profile in Base.gov. The full code is available in the support files.[11]

The scrapper for Parlamento.pt was built on two phases: First, based on selenium, the code simulates human navigation opening every interest register page. Selenium and a chrome web driver were essential because the Parlamento.pt website uses asp.net, meaning the content is generated at the moment and there are no permanent links to follow, so the solution is to simulate a click. When testing showed the scrapper was able to open every single interest register page and move on to the next, the second phase started.

Figure 6 shows the part of the code that identifies the number of registers per page and follows the link for each one of them in a loop until it reaches the last link. Then it repeats the process on the next page until it reaches the last link on the last page. Row 68 identifies the link by the text it contains, which is always the same, and row 69 is the click simulation.

[11] https://doi.org/10.34622/datarepositorium/KWFSU2

```
    # loop through every instance of a parliamentarian
    for each in parliamentariandiv[1:]:
        # name of person
        person = each.a.text

        # electoral cycle
        cycle = each.span.text

        # political party
        polParty = each.div.next_sibling.next_sibling.next_sibling.next_sibling.span.text

        # control prints to be transformed
        print("Nome: " + person)
        sleep(2)

        personalLink = each.find("a").get('href')


        # navigate into interests register incremental
        num = num + 1
        clicable = driver.find_elements_by_link_text('[ver...]')[num]
        clicable.click()
        sleep(5)
```

Figure 6 Parlamento.pt scrapper extract: Looping through Interest Registers

The second part of the code was written based on three downloaded pages with the interests register so it could be tested without sending multiple requests to the website. This essentially selects the needed data on each page, checks for blanks and identifies them, and appends the data to an array. This phase also includes the writing of scraped data in a CSV file with Pandas. The fact that every page follows the same structure made this process simpler and a matter of jumping between "if-else" conditional structures.

The resulting product is a CSV with one row per parliamentarian and one column per attribute. For attributes with multiple values, such as "Job Positions", where one person can have multiple job positions, the values are all in the same cell separated by two consecutive colons "::". Separating these values again is then handled in the data cleaning phase.

The data about interest registers in parlamento.pt is displayed as a form result. There is a title for current positions, then the information about the positions, then a title about old positions, and information about those. This, at times, leads to repeated information. For instance, if a parliamentarian works in a company since 2017, instead of the position showing only once as a current position that started in 2017, sometimes it appears, once as a past position "last three years", and once as a current one "accumulation". A translation by the author example is displayed in Table 17 Example of the information displayed on Parlamento.pt (city name withheld). The name of the city was removed to avoid singling out individuals.

Table 17 Example of the information displayed on Parlamento.pt (city name withheld)

III – Data About jobs/positions/activities	
Last Three Years	
Job/Position/Activity	Member of Local Parliament
Entity	Local Parliament of City 1
Start	2017 -10–22
End	
Accumulation with political/public chair position	
Job/Position/Activity	Member of Local Parliament
Entity	Local Parliament of City 1
Start	2017-10-22
End	

Although duplicate this information was kept, ensuring that important data was not accidentally excluded. When transforming the data sampled in Table 17 into a tabular format where there are four columns identifies whether: the job position, the entity, the start date, and the end date which have as values whichever value is on the website. Then, to avoid losing information, two columns were added, one to specify if the job is a social position, a working position, or a society, in Table 17 it is a working position because it is under "III – Data About jobs/positions/activities". The second column identifies whether the position happens in the accumulation with political work or not, in Table 17, that information is on lines two and seven.

Figure 7 shows how these two columns are added. Line 172 makes sure the data is only being retrieved in the section of the form with info on "social Positions", therefore, in line 189 the property "Social Position" is added to make sure this information is not lost. Then for the accumulation, in lines 190 to 196, the scrapper checks whether the information is under the accumulation separator. If yes, then the value is True, if not then the value for isAccumulation is false.

```
171         # IV - Social Positions
172         social = soup.find(id=re.compile('CargosSociais$'))
173
174         for activity in social.find_all(id=re.compile('Cargo$')):
175             activities.append(activity.text)
176         for entity in social.find_all(id=re.compile('Entidade$')):
177             entities.append(entity.text)
178         for domain in social.find_all(id=re.compile('AreaActividade$')):
179             domains.append(domain.text.replace(';', ':'))
180         for headquarter in social.find_all(id=re.compile('LocalSede$')):
181             headquarters.append(headquarter.text)
182         for startDate in social.find_all(id=re.compile('Cargo$')):
183             startdates.append('null')
184         for endDate in social.find_all(id=re.compile('Cargo$')):
185             enddates.append('null')
186         for percentage in social.find_all(id=re.compile('Cargo$')):
187             participation.append('null')
188         for activity in social.find_all(id=re.compile('Cargo$')):
189             workRelation.append('Social Position')
190         for accumulation in social.find_all(id=re.compile('Cargo$')):
191             control = accumulation.find_next('div', class_='row margin_h0 Titulo-Cinzento margin-Top-15 font20').text.strip()
192
193             if control == ('Acumulação com cargo político/alto cargo público'):
194                 isAccumulation.append('False')
195             else:
196                 isAccumulation.append('True')
197
198         # Control if all arrays have same lenght
199         if len(activities) == len(entities) == len(startdates) == len(enddates) == len(isAccumulation) == len(
200                 workRelation) == len(domains) == len(headquarters) == len(participation):
201             print('Activities Section IV is correct')
202         else:
203             print('verify data on activities section IV')
204
```

Figure 7 Parlamento.pt scrapper extract: Identifying Data on Social Positions

The last six lines in Figure 7 show the sum that is made to make sure the data is aligned and that every multivalued cell in one line has the same number of values.

At first sight, these additions seem redundant because one should be able to tell whether it is an accumulation by the start and end date, however a quick analysis of the registers showed that these dates are not always given so extracting as much context information from the page as possible was considered the better option.

Uniformize then proceeds with several data cleaning steps that result in a list of entities. This list then feeds a scrapper that looks for the entities in Base.Gov.

The scraper for Base.gov.pt was simpler to build. There is an array with company names where the spaces are substituted by "+" and one at a time they are substituted in the URL. Selenium is only used here for clicking the download button and checking the sum of contracts for control purposes.

For extracting the contracts from Base.gov.pt there were two options: Search the entity names as simple text, this would return contracts with the entity name in any of the contract's fields; Search for the entity name as a contractor and then as a supplier separately, this returns only contracts where the entity appears as contractor or supplier of the contract. The first option would return results faster, but the results were worst because if the name of the entity appeared in a field of the contract that was not the Supplier or the Contractor the contract would be downloaded anyway.

For instance, let's consider a company named "Creativity". A search for the text "Creativity" returns contracts where the supplier or contractor have the name "Creativity" and a contract from a public entity that contracted a professional school to give "A workshop to increase the creativity of the team". The last result is not of interest because the entity researched is not involved.

Considering the above, on phase one all entities are run in the "contractor" field three times, then in the "contracted" field three times and the ones that are not found in either are checked one last time on the "text" field. All 725 entities were searched as suppliers and contractors three times. The time the driver waited for an answer before moving to the next result increased in each attempt. After the third try, all entities that were not found were searched only once as free text.

All downloaded files were then verified manually to make sure the entities mentioned correspond to the entities on the Interests Registers. Whenever it was not possible to verify if it was the same entity due to lack of information the contracts from said entity were discarded. This process is described in Figure 8.

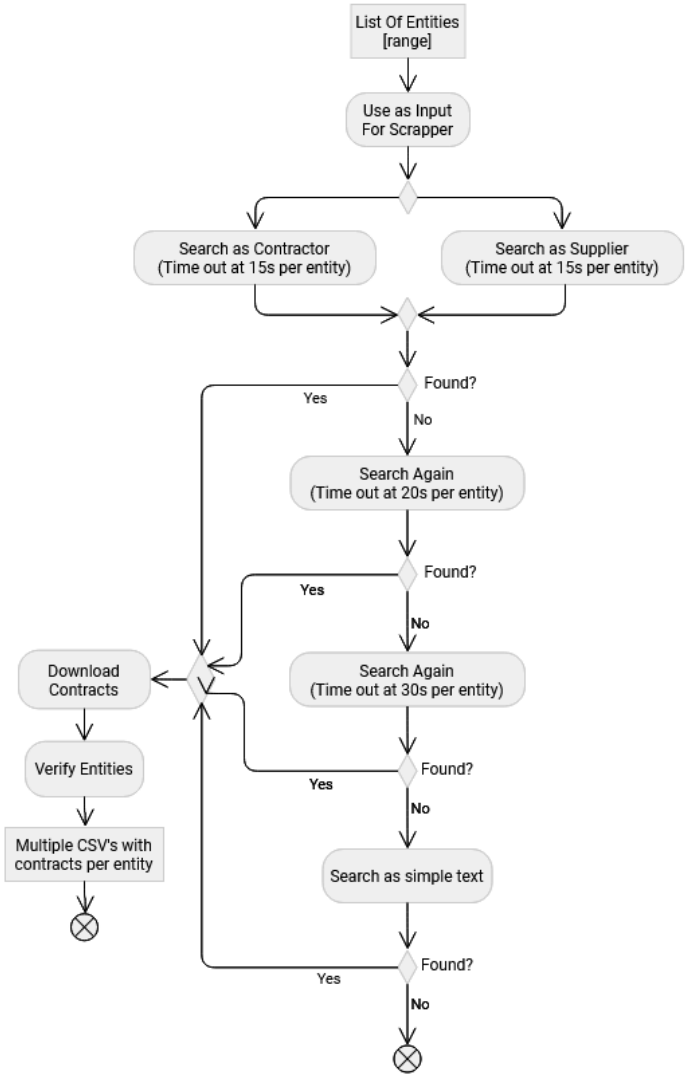


Figure 8 Retrieve Contracts

The data is then be treated on OpenRefine and a list of NIFs for all the entities already mentioned in the Parliamentarians data, plus the new ones from Base.gov.pt that interacted with them via contracts is extracted. This list of NIF's is searched on Base.gov.pt and the scrapper saves the link to their online profile on the website. This link is later used to enrich the data. -The code behind this process can be seen in Figure 9 and the output is a CSV with two columns, one with the NIF to cross with the remainder of the data and one with the link. They are identified in lines 31 and 32.

```
19  for page in pages:
20      nif = []
21      id = []
22      driver = webdriver.Chrome('C:/Users/Asus/Desktop/exercicios limpos/scrape/chromedriver.exe')
23      driver.get("https://www.base.gov.pt/Base4/pt/pesquisa/?type=entidades&texto=" + page)
24
25      sleep(25)
26
27      soup = BeautifulSoup(driver.page_source, 'lxml')
28      #all_links = soup.find('a', href=True)
29      div = soup.find('div', class_="b-resultados")
30
31      link = div.find_next('a', href=True)
32      URL = link.get('href')
33
34      print(page)
35      print(URL)
36
37      nif.append(page)
38      id.append(URL)
39
40      urls = pd.DataFrame({
41          'nif': nif,
42          'link': id
43      })
```

Figure 9 Base.gov.pt URL scrapper extract: Loop through NIFs and get Links

4.1.2 Cleaning and Uniformization

Figure 5 only represents the general steps that directly relate to the Data, but there is a lot of work behind each one of them. This chapter refers mainly to the decisions and steps behind the "Clean and Uniformize" activity nodes.

As represented in the same figure, the tool used here is OpenRefine which creates logs of every transformation that are available for consultation in the support files[12].

The cleaning process was carried out on four different files: Parliamentarians, Supplier Contracts, Contractor Contracts, and Entities. Since the process for both files with contracts was the same, from here on out both are simply referred to as Contracts. The only reason they were kept separate was that, as can be seen in Figure 8 they were sourced separately and kept that way during transformation to make tracing back issues easier.

[12] https://doi.org/10.34622/datarepositorium/KWFSU2

For the cleaning process, it is important to understand the size of the files. As stated in Table 18 there are 224 parliamentarians. None was eliminated, and even though the Portuguese parliament has 230 seats, at the time of retrieval only 224 interest registers were available. This 224 is counting the ones that do not have anything to declare but publish a document, nevertheless.

All 725 entities presented on the interest registers were searched on Base.gov.pt which returned 246 entities with contracts. 2721 Contracts have one of these 246 entities as a supplier and 148722 contracts have one of those entities as a contractor.

The Entities file is composed of all the entities found on Interest Registers (725), plus the entities representing the other side of the contracts retrieved from Base.gov.pt (24 305), giving us a total of 25 030 entities.

Table 18 Sum of Parliamentarians, Entities and Contracts

	Initial Sum	Final Sum
Parliamentarians	224	224
Contracts	148 722 + 2 721	129 011 + 2 656
Entities	-	25 519

No parliamentarians needed to be discarded, and the entities only have a final sum because the file was created from the other two after they were cleaned so no discarding was needed.

Discarded contracts include contracts that did not have at least one contractor and one supplier because this is the minimum needed information to describe a contract, per definition.[13] Moreover, to reduce complexity in this prototype contracts with multiple contractors or suppliers were also discarded since that would add a layer of complexity that was considered larger than the extra data it would provide. Eliminating contracts with multiple entities did not leave any of the 246 initially found entities without any contract.

4.1.2.4 Uniformization procedures:

Check for excessive and out of place data: Most of the data scraped from Parlamento.pt is typed in free text format. Meaning there are no pick lists or verification of data for things like addresses. Which leaves room for inconsistencies when referring to the same thing. The first step was getting rid of prefixes that add nothing to the data: in the column entity, a cell value "Parliamentarian in Portuguese Parliament" has excessive data. The entity is the Portuguese Parliament, the Parliamentarian is the role and belongs

[13] https://www.merriam-webster.com/dictionary/contract
Contract: a binding agreement between two or more persons or parties.

in another column. For the column entity, only entities are accepted. The inverse happens in the column role.

Check for typos and duplicates: Another common occurrence is mentioning the same entity or expression with different nomenclatures. The best example for this is "Portuguese Parliament" which in the original data is "Assembleia da República". This appears as "assembleia da República", "A.R.", "AR", "Parlamento", "assembleia republica" and other varieties. In every case found where it could be understood that the same entity was mentioned with different names, the most complete name was always chosen.

Various techniques can be used to look for similar entries on columns and fix issues such as the one with the varieties of names for the Portuguese Parliament. OpenRefine allows for comparison through clustering techniques such as neighbouring, fingerprint, anagrams and Levenshtein's distance. The same techniques were used by [49] and [40]. Every comparison was manually checked before making permanent edits to the data. Figure 10 shows a sample of the field "Reasons for changes in the contract's date" before uniformization. All the values visible in the print mean the same thing: "Not Applicable".

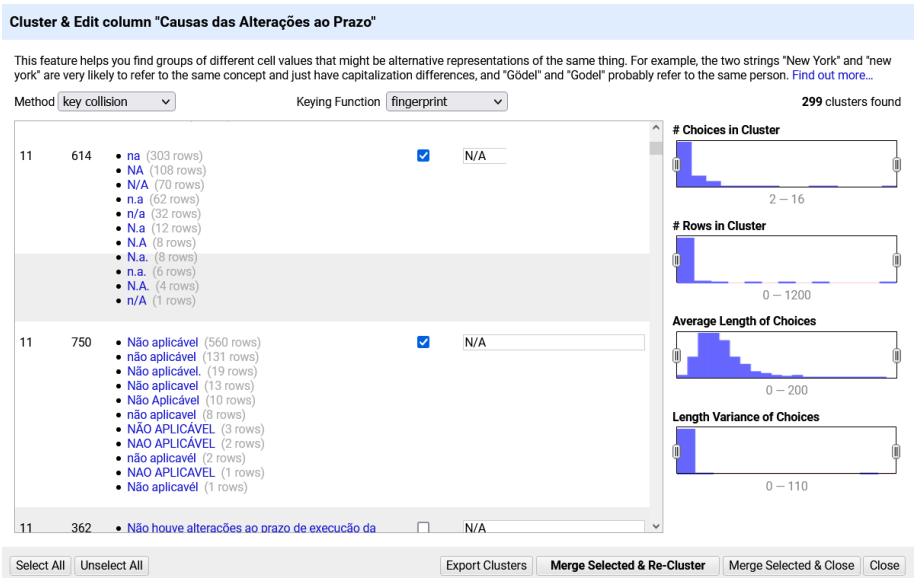


Figure 10 OpenRefine cleaning contracts: Clustering Techniques - Key collision – Fingerprint

This clustering process led to the reduction of unique values, but it was not enough, the remainder of the entities and roles were also corrected in other ways including: [14]

14 The data used in the examples is in Portuguese since it is taken from the original data. The stop words elimination and acronym extension are self-explanatory. The gender is not a translatable problem since in English the word lawyer ("advogado/a") is the same for both male and female. The synonyms in the example are the same as having "freelancer" and "individual service provider", they are interchangeable. Finally Unnecessary data here has the example of an answer to a job position that includes the place where the job is done. The place belongs to the organization field, not the position field so "Lawyer at A&B" would simply be "Lawyer".

- The elimination of stop words: ex.: “de”, “da”, “o”, “a”, “e”.
- Neutralization of gender in roles: ex.:“advogada” > “advogado”.
- Extension of acronyms: ex.: “A.R.” = “Assembleia da República”.
- Uniformization of obvious synonyms: ex.: “trabalhador por conta própria” = “Profissão liberal”.
- Simplification of unnecessary data: “Posição: Advogado na Firma A&B” = “Posição: Advogado”.

The data from the Contracts had better quality and for that reason, the transformation process was simpler and limited to the automatic clustering techniques mentioned above. These techniques were applied to the columns of the legal framework documents, the causes for alteration in price, and the causes for alterations in calendarization.

For the entities in contracts with the same NIF and slightly different names, the most complete name per NIF was chosen. When the names were different names for the same entity an “Alternative Names” column was created. An example of this is the research centres of universities. Their fiscal number is the same as the parent institution, but they have a different name, so the name was kept as “Alternative Name”. In this dataset, the political parties with a seat on parliament were also manually added as entities. In the properties for disclosing motives for price and date modification a lot of strings with variations of “Nothing changed” represent excessive data that adds no information.

Cleaning the original data was the first big challenge encountered in the implementation phase. There were both excessive and lacking data in the contracts and the interest registers. Automating this cleaning process would be very time-consuming because the data is not uniform and for that reason, a lot of the cleaning had to be done manually.

The last step on every dataset for the parliamentarians, the entities, and the contracts were creating an internal IRI for linking the data. The local IRI for each parliamentarian is a “P” plus their parliamentarian code from Parlamento.pt. For the organizations with a profile on Base.Gov, it’s an “O” plus their profile number from Base.Gov. For the 498 organizations mentioned in the interest registers and not found in base.gov, the IRI is “OL” plus their row number. The contracts are identified by a “C” plus the NIF of the contractor, plus the NIF of the supplier, plus the row number to avoid the repetition of IRI’s.

4.2 Designing the Linked Data App Profile

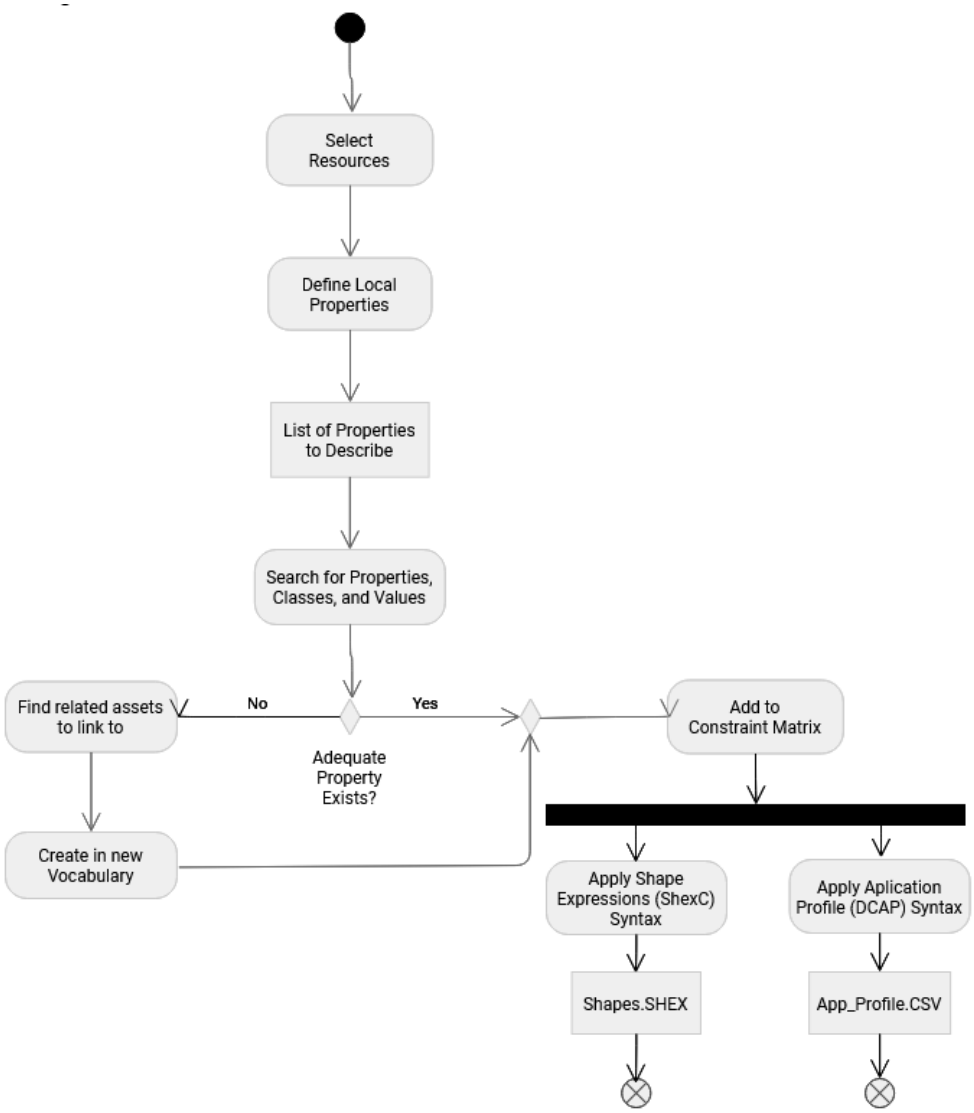


Figure 11 Design Linked Data Profile

The second large activity introduced above in Figure 4 Data Journey - From Sourcing to LOD is broken down into finer steps above in Figure 11. This chapter goes on about the planning and designing work that should be done before starting the transformation of the tabular data into LOD. The starting point is the properties from Table 14. The first step in designing the LOD Application Profile is structured using a constraint matrix.

A Constraint Matrix is essentially a Class-based table that describes the LOD properties used for each subject based on their class. The first column has the name of the property like in Table 14, then the following columns have the link to the appropriate existing property, the description, the original domain, the original range, the allowed values, and the cardinality. This table was populated by following the process described in Figure 11 Design Linked Data Profile. The final Constraint Matrix is available in

Appendix C where it shows onto what types of resources each property is used, with what class of values they related to, and how the resources relate to each other.

This matrix is based on the DCTAP application profile but has extra information to help on an initial phase. This refers particularly to the original range and domain that help avoid vocabulary violations. Below is the description of the process of choosing and allocating properties and classes.

The first approach to describe the data was trying to use only already existing vocabularies prioritizing common and popular ones. [27] Previous research had already indicated three big ontologies that seemed perfect for describing the data, they are described in Table 19.

Table 19 First Ontologies used.

Ontology	**The Italian Open Parliament initiative**[15]	**Public Procurement Ontology**[16]	**The Organization Ontology**[17]
Description	This ontology describes parliamentarians and makes all LOD descriptions available to the public on their website.	PPROC is a project aiming to add semantic technologies to describe procurement procedures. This is focused on Spanish and European laws.	A core ontology for an organization supports a range of domains inside the organization framework.
Pros	Has an in-depth description of parliamentarians Is used by LinkedEP [43] to link their properties.	The ontology is very complete and nearly everything applies since the whole procedure is very similar to what happens in Portugal. It is well connected to strong ontologies.	It is possible to map the roles and positions of the parliamentarians in organizations without the use of blank nodes or reification.
Cons	The properties are in Italian. They do not follow the convention of capitalizing classes to distinguish from properties. There are a lot of constraints that would lead to redundant data.	The ontology is very precise, and the constraints placed on the ranges and domains make it very complex for the data at hand. There would be multiple single property subjects.	There is not enough information in the data to justify the use of Membership and Role as subjects.

The first attempt consisted in using only these ontologies to describe as many attributes from Table 14 as possible with properties and classes from these three ontologies and looking in other vocabularies to fill in the missing attributes. This process led to a constraint matrix with over 60 properties (seven would be new), 15 namespaces, and 11 described subject types, out of which 7 had less than 3 properties.

15 https://data.camera.it/data/en/datasets/
16 http://contsem.unizar.es/def/sector-publico/pproc.html
17 https://www.w3.org/TR/vocab-org/

In this matrix, the parliamentarian is linked to the entity through the subject membership. This linking option did was discarded for the following reasons:

- Parliamentarians are not directly connected to the entity. They are only connected to the entity on a second level by the place of the membership.
- Obeying the domains and ranges creates too many resources to describe and there is not enough information for this to be reasonable since it creates several subjects with less than three predicates.
- Having several levels of data creates more namespaces and more complex queries. This impacts query response time in the future and generates a heavier dataset.

The idea of merging all these vocabularies would have been ideal if there was control over the information and the scope of the data sources was larger. What ended up happening was that by having three classes of resources on the dataset and having only first level information about each of them, representing the data on multiple levels adds more complexity and heavier datasets but does not add more information.

From this experience the process re-started with three goals: give priority to well-established property schemas, having a property with a parliamentarian as subject and an entity as an object or an entity as object and parliamentarian as subject, and using fewer vocabulary namespaces.

This second attempt uses only 50 properties (6 had to be created), 9 namespaces, 5 described subjects and 2 blank nodes. No data was discarded, the reduction of properties is due to the reduction of subjects caused by the choice of less restrictive properties.

Selecting the properties to re-use, defining the shapes of the data, and trying to avoid excessive blank nodes constituted one of the biggest challenges of the implementation phase. There are many ontologies in the LOD universe, but when one tries to take from different ontologies to create a multidisciplinary database, obeying the ranges and domains becomes a considerable challenge. With the Public Procurement and Public Contracts Ontology, this was the case because the properties are very restrictive. The description and the names of the properties were adequate to every situation but then the data available was not enough to justify diving into sublevels of subjects for the price, legal documents, or framework agreements. To make querying more straightforward, and to keep the number of blank nodes as low as possible in the prototype, the final model opts for having as many properties on the first level as possible.

To aid in visualizing the properties mentioned, and how the subjects connect to one another, an adaptation of an Entity Relational model was elaborated and can be seen below in Figure 12 Adapted Linked Data Entity Relation Model. This Image does not follow the specifications for an ER model diagram,

the correspondences are that each table is a subject. The class of the subject is the title of the table, inside each table the left side has the properties and on the right side, the value types or controlled vocabulary stems. The arrows connecting the tables point to the value: A pc:Contract has as pc:supplier a schema:Organization. For the values with vocabularies, skos:Concept represents the class of the value, followed by the suffix of said vocabulary and the "~" is a placeholder meaning that any value that starts with that root applies.

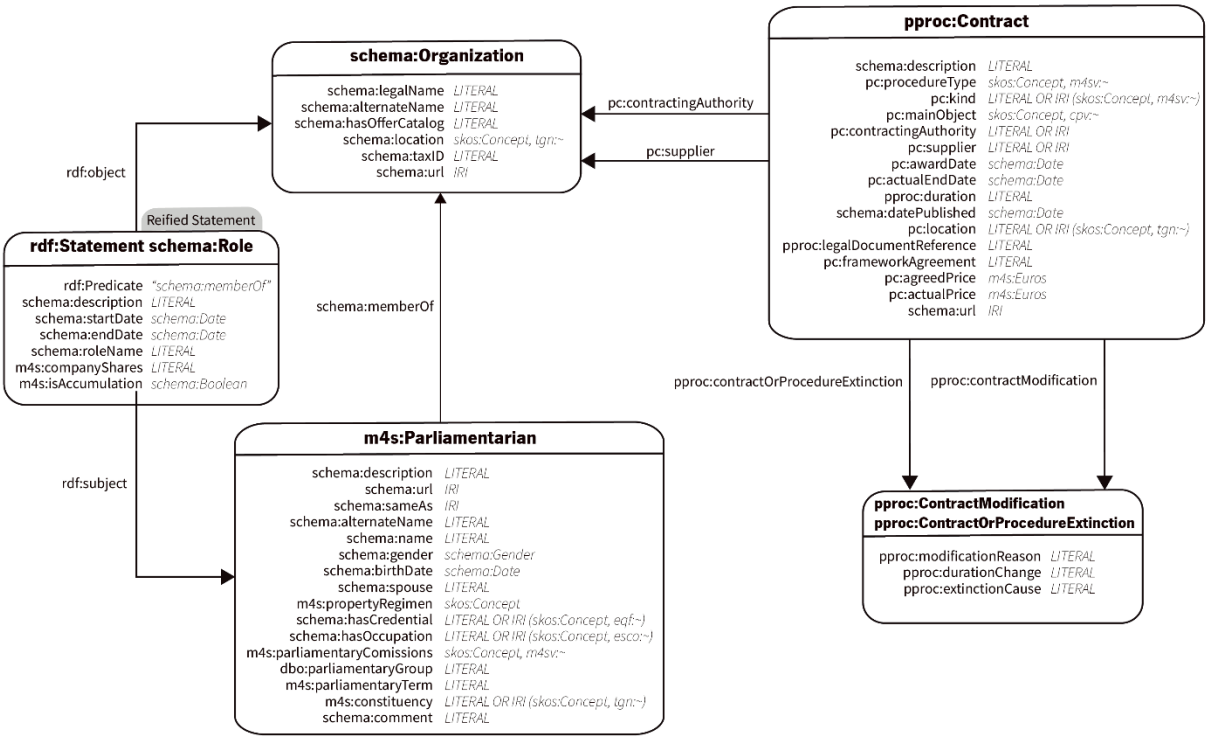


Figure 12 Adapted Linked Data Entity Relation Model

The main vocabulary of properties used for describing what is seen in Figure 12 is Schema.org. This is a well-maintained project with community participation and an up-to-date GitHub for support and discussions. Moreover, it was created as a joint effort from Google, Microsoft, and Yahoo! as a markup language to be read by search engines that would use it to provide richer results. [4]

Schema.org is used on 23 properties and replaced every instance of the organization ontology that was used on the first matrix.

4.2.1 Reification

In Figure 12 the table further to the left has a note on top saying "Reified Statement". This chapter explains why it is needed and how its format was selected.

Besides the creation of new properties and vocabularies, one challenge was the representation of the connection between person and entity with reification. The first is discussed in the next chapter, the latter is explained below.

Let us consider the following information as an example:

The parliamentarian "John Doe" had a Role of consultant in "Society A" until 2019.

If it were not for the year and the role consultant, the representation of this as a triple would be simple:

John Doe – Subject > Member of – Predicate > Society A – Object

However, both the data about the year and the position are important. This is a case of handling knowledge about knowledge which is known as metaknowledge [62] and can be represented in several ways in RDF.

Metaknowledge usually refers to information about data provenance, reliability, and timestamp. However, understanding its usability is complex [62] and the same techniques used to describe it are also used for describing different data. One of the biggest and most common examples are the qualifiers in Wikidata. Wikidata has qualifiers on most resources, they have a large collection of resources, and they include politicians, even some politicians from the Portuguese parliament. The way they organized the information about parliamentarians served as a base to organize the application profile, (even though their properties were avoided for lack of human readability), especially the use of qualifiers to add context, and extra information about statements. Wikidata offers an RDF dump which shows that the syntax they use in RDF for describing their qualifier is a form of reification that does not follow the standard RDF reification. The RDF standard for reification is sometimes controversial causing the appearance of uses like this that are inspired in the norm, use the type "rdf:Statement" but don't follow the guidelines. [63]

Wikidata introduced the RDF dump in April 2015, before that, both [64] and [63] discussed the representation of Wikidata's syntax in RDF and went through different reification options. The first compared four ways of achieving the result: standard RDF reification, n-ary relations, singleton properties and named graphs. They tested all four versions by querying them with SPARQL and uploading them to different Triplestore and concluded that there was no clear winner in terms of query performance. However, some Triplestores did struggle with processing singleton properties. [64] The second ended up narrowing it down to standard RDF reification (without obeying the formal guidelines) or named graphs as the best options for representing the qualifiers. [63]

Below in Table 20 is an example of what the triple in the example above looks like as a Wikidata qualifier, as a named graph, and as standard RDF reification.

Table 20 Reification samples Wikipedia Named Graphs and Standard Reification

Wikidata Qualifiers	**Named Graphs**	**Standard RDF Reification**
:JohnDoe :hasRole :Consultant :JohnDoe :hasRole :S1 . :S1 :hasRole :Consultant; :S1 :ended "2019"; :S1 :entity :SocietyA .	:JohnDoe :hasRole :Consultant :Graph1 . :Graph1 :entity :SocietyA; :Graph1 :ended "2019" .	:JohnDoe :hasRole :Consultant . S1 rdf:type rdf:Statement S1 rdf:object :John doe S1 rdf:predicate :hasRole S1 rdf:Object :Consultant S1 :ended "2019" S1 :entity :SocietyA

More recently there is also the work of [65] that mapped a relation between Wikidata and DBpedia. This included relating the properties with "owl:equivalentProperty" as well as the information. When facing the issue of reification, they make an important note of the fact that DBpedia has a more stable ontology than Wikidata that was not native RDF. This is one of the reasons why they select simple RDF reification as the go-to technique for mapping Wikidata with DBpedia. [65] Both for the sake of scientific work and the reputation of DBpedia in the world of linked data this prototype is going to use standard RDF reification. Literature shows it works, it has solid guidelines and a standard way to be described[18], and the most widely discussed downside, which is the number of triples it generates does not have as big an impact as one would expect when it comes to query time response. [64] Named graphs also seemed like an adequate solution and a largely supported one, however, OpenRefine is selected as the tool to transform the data into RDF and that also presents a constraint since a way to create named graphs in OpenRefine was not found. Figure 13 shows what a standard reification mapping looks like in OpenRefine and what the outputs looks like in Turtle. Querying reified data will be discussed further down in Publication.

[18] https://www.w3.org/TR/rdf-primer/#reification

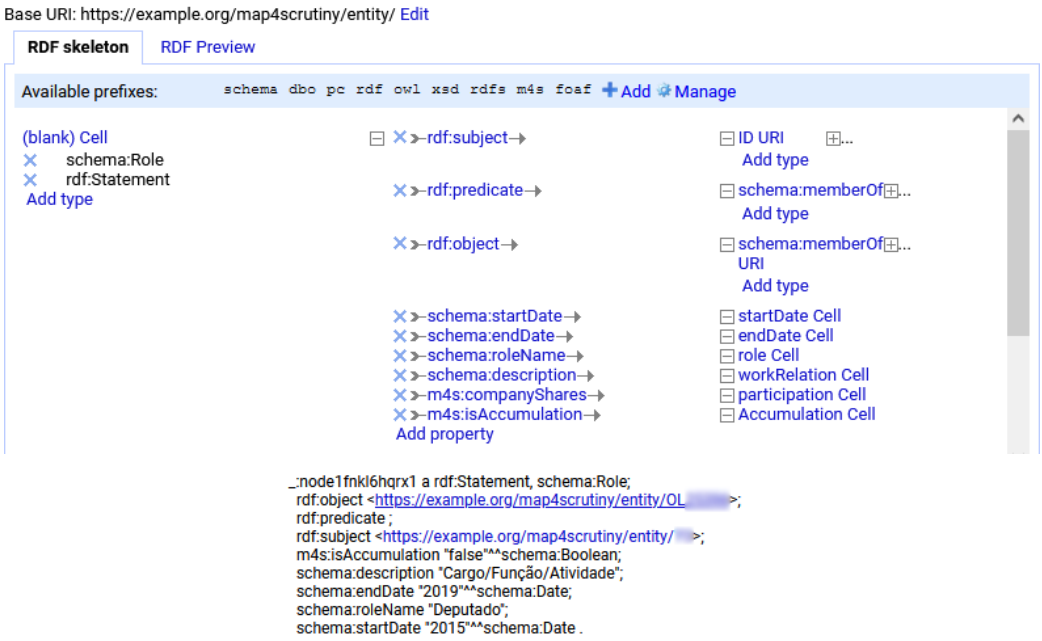


Figure 13 OpenRefine: Reification Map and Turtle Sample (identifiers blurred)

4.2.2 Dublin Core Application Profile (DCTAP)

The DCTAP , presented above in 2.3.2 Linked Data, being an ongoing project naturally means it is not finished, however, their tabular approach is already very comprehensive and was, therefore, used to describe the application profile for this project as can be seen in Table 21.

The template for the table as well as the guidelines followed are all available on their GitHub. [19] The representation in Table 21 is only missing the label for the shapes since they are self-explanatory and the label for the properties that can be found in Table 14.

This format is preferred to the Matrix and the ER adaptation mentioned above because it is a format with guidelines for implementation available to the public, meaning that it can be interpreted by anyone. The other two are working documents that aid in planning and visualizing but are not supported by research as rich as this version.

[19] https://github.com/dcmi/dctap/blob/main/TAPprimer.md

Table 21 Application Profile - DCTAP Template

Prefix	Namespaces
dbo:	http://dbpedia.org/ontology/
pc:	http://purl.org/procurement/public-contracts#
pproc:	http://contsem.unizar.es/def/sector-publico/pproc#
rdf:	http://www.w3.org/1999/02/22-rdf-syntax-ns#
schema:	http://schema.org/
m4s:	<www.example.org/map4scrutiny/core#>
m4sv:	<www.example.org/map4scrutiny/value#>

shapeID	propertyID	mandatory	repeatable	Value Node Type	Value Data Type	valueConstraint	valueShape	note
Parliamentarian								
	rdf:type	TRUE	TRUE	IRI		m4s:Parliamentarian		
	schema:url	TRUE	FALSE	IRI	IRIStem	www.Parlamento.pt/DeputadoGP/Paginas/Biografia.aspx?BID=~		
	schema:sameAs	FALSE	TRUE	IRI				
	schema:description	TRUE	FALSE	LITERAL				
	schema:alternateName	TRUE	FALSE	LITERAL				
	schema:name	TRUE	FALSE	LITERAL				
	schema:gender	TRUE	FALSE	IRI	IRIStem	schema:GenderType		
	schema:birthDate	TRUE	FALSE	LITERAL	schema:Date			ISO 8601 date format.
	schema:spouse	FALSE	FALSE	LITERAL				
	m4s:propertyRegimen	FALSE	FALSE	IRI	IRIStem	m4sv:PropertyRegimen		
	schema:hasCredential	FALSE	TRUE	IRI	IRIStem	http://data.europa.eu/snb/isced-f/		
	schema:hasOccupation	TRUE	FALSE	IRI	IRIStem	http://data.europa.eu/esco/isco/		
	m4s:parliamentaryComissions	FALSE	TRUE	IRI	IRIStem	m4sv:ComissoesParlamentares		
	dbo:parliamentarygroup	FALSE	TRUE	LITERAL				
	schema:memberOf	TRUE	TRUE	IRI			Organization	
	m4s:ParliamentaryTerm	TRUE	TRUE	LITERAL				roman numbers
	m4s:constituency	TRUE	TRUE	IRI	IRIStem	http://vocab.getty.edu/tgn/		
	schema:comment	FALSE	FALSE	LITERAL				

shapeID	propertyID	mandatory / repeatable		Value Node Type	Value Data Type	valueConstraint	valueShape	note
_:Roles								
	rdf:type	TRUE	TRUE	IRI		rdf:Statement		
	rdf:type	TRUE	TRUE	IRI		schema:Role		
	rdf:subject	TRUE	FALSE	IRI			Parliamentarian	
	rdf:predicate	TRUE	FALSE	IRI		schema:memberOf		
	rdf:object	TRUE	FALSE	IRI			Organization	
	schema:roleName	TRUE	FALSE	LITERAL	IRIStem	http://data.europa.eu/esco/isco/		
	schema:endDate	FALSE	FALSE	LITERAL	Schema:Date			
	schema:startDate	FALSE	FALSE	LITERAL	Schema:Date			
	schema:description	FALSE	FALSE	LITERAL				
	m4s:company Shares	FALSE	FALSE	LITERAL				
	m4s:is Accumulation	FALSE	FALSE	LITERAL	Schema:Boolean			

shapeID	propertyID	mandatory / repeatable		Value Node Type	Value Data Type	valueConstraint	valueShape	note
Organization								
	rdf:type	TRUE	TRUE	IRI	IRIStem	schema:Organization		
	schema:legalName	TRUE	FALSE	LITERAL				
	schema:alternate Name	FALSE	TRUE					
	schema:hasOffer Catalog	FALSE	FALSE	LITERAL				
	schema:location	FALSE	TRUE	IRI	IRIStem	http://vocab.getty.edu/tgn/		
	schema:taxID	FALSE	FALSE	LITERAL				
	schema:url	FALSE	FALSE	IRI				

shapeID	propertyID	mandatory	repeatable	Value Node Type	Value Data Type	valueConstraint	valueShape	note
Contract								
	rdf:type	TRUE	TRUE	IRI		pproc:Contract		
	schema:description	TRUE	FALSE	LITERAL				
	pc:procedureType	FALSE	TRUE	IRI	IRIStem	m4sv:TipoProcedimento		
	pc:kind	FALSE	TRUE	IRI	IRIStem	m4sv:TipoContrato		
	pc:mainObject	TRUE	FALSE	IRI	IRIStem	http://purl.org/cpv/2008/		
	pc:contracting Authority	TRUE	FALSE	IRI			Organization	
	pc:supplier	TRUE	FALSE	IRI			Organization	
	pc:awardDate	FALSE	FALSE	LITERAL	Schema: Date			
	pc:actualEndDate	FALSE	FALSE	LITERAL	Schema: Date			
	pproc:duration	FALSE	FALSE	LITERAL				
	schema:datePublished	FALSE	FALSE	LITERAL	Schema: Date			
	pc:location	FALSE	TRUE	IRI	IRIStem	http://vocab.getty.edu/tgn/		
	pproc:legal Document Reference	FALSE	FALSE	LITERAL				
	pc:framework Agreement	FALSE	FALSE	LITERAL				number and description
	pc:agreedPrice	FALSE	FALSE	LITERAL	m4s:Euros			in euros
	pc:actualPrice	FALSE	FALSE	LITERAL	m4s:Euros			in euros
	schema:url	FALSE	FALSE	IRI				
	pproc:contract Modification	FALSE	FALSE	bNode			_:Cmodification	
	pproc:contractOr Procedure Extinction	FALSE	FALSE	bNode			_:Cmodification	

shapeID	propertyID	mandatory	repeatable	Value Node Type	Value Data Type	valueConstraint	valueShape	note
_:Cmodification								
	rdf:type	FALSE	FALSE		IRIStem	pproc:ContractModification		
	rdf:type	FALSE	FALSE		IRIStem	pproc:ContractOr ProcedureExtinction		
	pproc:modification Reason	FALSE	FALSE	LITERAL				
	pproc:duration Change	FALSE	FALSE	LITERAL				
	pproc:extinctionCause	FALSE	FALSE	LITERAL				

4.2.3 Vocabulary Creation

Everything in Table 21 with the prefix (m4s:) represents a vocabulary that is going to be created. This chapter explains what led to this decision and, when appliable, to what other properties the new ones are being linked to, and why these properties were not used in the first place.

Table 22 Created Properties and Classes

Object Properties	**Linked to**	**Motive**
propertyRegimen	schema:spouse	The property regimen is important for the data and connected to the marriage, but no property was found to represent this. For this reason, a new one is being created as a sub-property of "schema:spouse" since it is a further specification of the marriage.
constituency	wiki:P768(electoral district)[a] admingeo:constituency	The only options found for this property are not ideal and for this reason, the decision was to create a new one that has a "same as" connection to the ones mentioned.
parliamentaryComissions	ocd:rif_ufficioParlamentare[b]	The only property found has incompatible range and domain being used as a same as instead of simply reused.
parliamentaryTerm	ocd:rif_leg[b] wiki:P2937 (parliamentary term)[a]	The only options found for this property are not ideal and for this reason, the decision was to create a new one that has a "same as" connection to the ones mentioned.
Datatype Properties	**Linked to**	**Motive**
companyShares	schema:owns	No property was found to represent this so schema:owns was selected as a super property since having shares is a type of ownership.
isAccumulation	-	This property is a Boolean describing whether the role it is used on happened in accumulation with public office. It is needed because start and end dates are not always specified.
Classes	**Linked to**	**Motive**
Parliamentarian	ocd:deputato[b] schema:Person	It is a subclass of schema:Person and the same as the OCD class.

The properties in the columns "linked to" were excluded for one of the following reasons:

a – Using codes for coding properties jeopardizes human readability.

b – These resources are only described in Italian and do not follow the norm of capitalizing classes and using lowercase on properties.

Admingeo:constituency was discarded because the description specifies a country.

Every re-used property is annotated in the ontology with a translation to Portuguese since that is the language of the original data, a schema:definedBy property with the link to the original description as a

value, and, when necessary, an addition to the domain and range. Extensions of the domain and range were only applied when the property made perfect sense, but the original range was for instance an IRI and most objects in the dataset are IRI's, but some are strings, or when a datatype was more suitable. The properties used for this purpose were schema:rangeIncludes and schema:domainIncludes. These properties were also used to define the range and domain of the new properties and classes. This way, the intended use is clear, and third parties have the basis to interpret whether the assets are appropriate. This solution is preferred because the description of an asset can suit a different, but very similar, range or domain than the one mentioned that makes more sense in a different implementation.

For the same reason, despite being identified as "Object" and "Datatype" properties in Table 22 in the ontology, the new properties have as rdf:type only rdf:Property to avoid excessive restraints. The ontology also includes a new datatype named "Euros" to be used in the price spaces to identify the currency.

```
:parliamentarianCommissions
    a                       rdf:Property;
    rdfs:label              "Work Groups and Commissions within parliament"@en, "Grupos e Comissões Parlamentares"@pt;
    schema:domainIncludes   schema:Person;
    schema:rangeIncludes    m4s:ComissoesParlamentares,
                            schema:Text;
    owl:equivalentProperty  <http://dati.camera.it/ocd/reference_document/#rif_ufficioParlamentare>
    .

:propertyRegimen
    a                       rdf:Property;
    rdfs:label              "Matrimonial Property Regime"@en, "Regime de Propriedade"@pt;
    schema:domainIncludes   schema:Person;
    schema:rangeIncludes    m4s:RegimePropriedade,
                            schema:Text;
    rdf:seeAlso             <https://schema.org/spouse>
    .

:companyShares
    a                       rdf:Property;
    rdfs:label              "Shares owned by shareholder"@en, "Ações detidas por um Acionista"@pt;
    rdfs:comment            "Can be represented as the monetary value of the actions, percentage or another value."@en,
                            "Pode ser representado através do valor total das ações, a quantidade ações ou outro valor."@pt;
    schema:domainIncludes   schema:Thing;
    schema:rangeIncludes    schema:Text;
    rdfs:subPropertyOf      schema:owns
    .
```

Figure 14 Extract from the Map4Scrutiny Ontology with three new properties

Figure 14 shows the implementation of Table 22 with triples for three of the new properties all linked to external vocabularies in different ways. ''Parliamentarian commissions" has an equivalent property in an external vocabulary, the "Property Regimen" does not have a super property nor an equivalent one, however, the concept is semantically related to that of a spouse in the sense that an application profile needing this property, probably also needs schema:spouse, thus the rdfs:seeAlso. The case for "Company shares" is different, holding actions from a society is a type of ownership, and for that reason schema:owns is an appropriate super-property. The full ontology with new and re-used assets is available in the support files[20].

[20] https://doi.org/10.34622/datarepositorium/KWFSU2

Table 23 Created Value Vocabularies

Controlled Vocabulary	Sources	Motive
ContractTypeVoc	Base.gov	There are only seven contract types, and the contracts always belong to one of them enforcing the sense of describing the options.
ParliamentaryComissionsVoc	Parlamento.pt	The commissions have goals and there is also a limited number of them. There are 44 including workgroups.
ProcedureTypeVoc	Base.gov	There are 13 procedure types, and the contracts are limited to these.
PropertyRegimenVoc	Eportugal.gov.pt	There are only 3 concepts for civil status recognized by the Portuguese civil code: Common Property, Separate Assets, and Community Property. There is also an "other" to leave space for marriages consolidated abroad.
ESCO Occupations	European Commission	This vocabulary is available to download in CSV format and via API in JSON. For this project, a transformation of the tabular vocabulary to SKOS was done. This vocabulary was not created, only transformed.

As for the controlled vocabularies, the decision process was simpler. First, there was a search for controlled vocabularies of terms on every Object that seemed appropriate. This search had two phases, first looking for the vocabularies in already known sources such as the European Union Publications Office and the United Nation Thesaurus. These sources are considered known because they were mentioned in the literature. The second phase included a free keyword search with standard web search engines for appropriate vocabulary. When the Object possible values suited the creation of a controlled vocabulary, but nothing was found the decision was to create a new one.

This application profile links to external value vocabularies for the CPV - the common procurement vocabulary, the TGN – taxonomy for geographical names, and the schema gender vocabulary.

```
:ContratosPublicos a    skos:ConceptScheme;
  skos:prefLabel        "Termos para Descrição de Contratos Públicos"@pt;
  dct:creator           "Inês Lopes";
  dct:contributor       "Ana Alice Batista", "Óscar João Atanázio Afonso";
  dct:created           "2021-07-05"^^schema:Date;
  dct:source            <https://data.dre.pt/eli/dec-lei/18/2008/p/cons/20210521/pt/html>;
  skos:editoralNote     "Usado na descrição de recursos classificados como: <http://contsem.unizar.es/def/sector-publico/pproc#Contract> no âmbito legal português."@pt;
  rdf:seeAlso           <http://purl.org/cpv/2008/>;
  skos:hasTopConcept    :TipoContrato, :TipoProcedimento .

:TipoContrato a      skos:Concept;
  skos:prefLabel      "Tipos de Contratos Públicos"@pt;
  skos:definition     "Tipos de Contractos definidos nos artigos do Código da Contratação Pública e usados no portal <base.gov.pt>"@pt;
  skos:editoralNote   "Narrowers são usados para descrever valores da propriedade:  <http://purl.org/procurement/public-contracts#kind> no âmbito legal português.>"@pt;
  skos:inScheme       <https://example.org/map4scrutiny/value#ContratosPublicos>;
  rdf:seeAlso         <https://raw.githubusercontent.com/opendatacz/public-contracts-ontology/master/schemes/contract-kinds.ttl>;
  skos:narrower       :BensMoveis, :Servicos, :ConcessaoObrasPublicas, :ConcessaoServicosPublicos,
                      :Sociedade, :EmpreitadasObrasPublicas, :LocacaoBensMoveis .

:TipoProcedimento a      skos:Concept;
  skos:prefLabel      "Tipos de Contratos Públicos"@pt;
  skos:editoralNote   "Narrowers são usados para descrever valores da propriedade: <http://purl.org/procurement/public-contracts#procedureType> no âmbito legal portugué
  skos:inScheme       <https://example.org/map4scrutiny/value#ContratosPublicos>;
  rdf:seeAlso         <https://raw.githubusercontent.com/opendatacz/public-contracts-ontology/master/schemes/procedure-types.ttl>;
  skos:narrower       :AjusteDiretoRG, :ConsultaPrevia, :ConcursoPublico, :ConcursoLimitado, :ProcedimentoNegociacao, :DialogoConcorrencial,
                      :AcordoQuadro258, :AcordoQuadro259, :ParceriaInovacao, :DispBensMoveis, :Serviços, :ConcursoConcecao, :ConcursoIdeias .

:BensMoveis a skos:Concept;
  skos:prefLabel      "Aquisição de Bens Móveis"@pt;
  skos:definition       "Entende-se por aquisição de bens móveis o contrato pelo qual um contraente público compra bens móveis a um fornecedor. (Artigo 437º CCP)"@pt;
  skos:broader        :TipoContrato .

:Servicos a skos:Concept;
  skos:prefLabel      "Aquisição de serviços"@pt;
  skos:definition     "Entende-se por aquisição de serviços o contrato pelo qual um contratante público adquire a prestação de um ou vários tipos de serviços mediante c
  skos:broader        :TipoContrato;
  skos:closeMatch     <http://purl.org/procurement/public-contracts-kinds#Services> .
```

Figure 15 Extract from the Map4Scrutiny Vocabulary of Values with Concept Scheme, Top Concepts, and Narrower Concepts

Figure 15 shows the conceptual schema for vocabularies used in the contract's procedure types and kinds. For both these top concepts, an approach already existed, it's linked in rdfs:seeAlso. However, the terms are not the same, and for legal reasons, only the obviously similar concepts were identified with skos:closeMatch as is the case of "Serviços" and "Services"

For the sake of consistency and incentive to re-use the new properties, class, and controlled vocabularies are described in RDF and SKOS based on the description of the resources being reused in this project. This means using the RDF primer as a guideline and the code from the namespaces as an example for the description of properties and the class. For the controlled vocabularies, this means basing the work on the syntax used by EU controlled vocabularies, which were also reused by other projects [48].

4.2.4 Shape Expressions

The ShEx schema equivalent to the DCTAP in Table 21 was written manually with ShExC. A compact and very human-readable form of the language for describing Shapes Expressions whose syntax is based on Turtle. The shapes have all the information presented above in the application profile and are available for consultation online in the support files[21] and Appendix E Shape Expressions. In this situation ShExC was chosen because the syntax is broad enough to describe everything needed in the application profile and because, coming from a DCTAP tabular profile, which is based on ShEx, this would be a natural choice. [46] The Shapes have two main uses: to aid a future user in finding his way through the triples by knowing how they are shaped and to be used for validation at a later stage. Some samples of data profiling that is described with ShEx with more detail than with the DCTAP are the values in which one can have either a vocabulary or a string, and the difference between open and closed shapes. In the first case, one can specify that a cell can have an xsd:string or a vocabulary with an IRIstem, but there is no way to specify that it must have both, or that it must have a string and may or may not have a member of a vocabulary. As for the open and closed shapes, in ShEx it is possible to classify a shape as such being the difference that an open shape must comply with the shape described but can have additional properties, while a closed shape must have only the properties predicted in the Shape Expressions.

This comparison is showcased in

Figure 16. The upper half is the application profile for the _:Roles and the bottom half the corresponding shape in ShExC. The content of the left box is almost equal in both formats, the substitution of rdf:type for "a" is a feature that ShExC borrowed from Turtle but using "a" and rdf:type is the same thing. The

[21] https://doi.org/10.34622/datarepositorium/KWFSU2

difference here is the “BNODE EXTRA a”, bNode defines that it is a blank node, and the Extra is the keyword that defines whether the shape is open or closed. In this case, it is open. This is not yet possible to define in DCTAP.

The middle section has the most differences in syntax, and essentially the info from the four columns is all there. The [ ] are used for specific values, the addition of ‘~’ in the end means that it is a prefix and is very useful for vocabulary. The validator will accept any IRI that starts with the given root or a literal but it won’t accept an IRI different from the one specified. The last line is the one with the most differences. ShEx automatically validates xsd:boolean, but the same is not true for schema:Boolean. This means that if the value type is xsd:boolean ShEx will only accept as values true and false written in lowercase without parentheses and everything else will be nonconformant. But if the value type is schema:Boolean and the value in the turtle file is for instance “not true”^^schema:Boolean ShExC will accept it and consider it conformant.

To avoid that possibility the two possible values with the identified datatype were specifically defined.

shapeID	*propertyID*	*ValueNode Type*	*ValueDataType*	*valueConstraint*	*valueShape*	*mandatory*	*repeatable*
_:Roles							
	rdf:type	IRI		rdf:Statement		TRUE	TRUE
	rdf:type	IRI		schema:Role		TRUE	TRUE
	rdf:subject	IRI			Parliamentarian	TRUE	FALSE
	rdf:predicate	IRI		schema:memberOf		TRUE	FALSE
	rdf:object	IRI			Organization	TRUE	FALSE
	schema:roleName	LITERAL		http://data.europa.eu/esco/isco/		TRUE	FALSE
	schema:endDate	LITERAL	Schema:Date			FALSE	FALSE
	schema:startDate	LITERAL	Schema:Date			FALSE	FALSE
	schema:description	LITERAL				FALSE	FALSE
	m4s:companyShares	LITERAL				FALSE	FALSE
	m4s:is Accumulation	LITERAL	Schema:Boolean			FALSE	FALSE

```
:metaRole BNODE EXTRA a {
  a                       [rdf:Statement]                                  ;
  a                       [schema:Role]                                    ;
  rdf:subject              IRI                                             ;
  rdf:predicate           [schema:memberOf]                                ;
  rdf:object               IRI                                             ;
  schema:description       LITERAL                                       ? ;
  schema:startDate         schema:Date                                   ? ;
  schema:endDate           schema:Date                                   ? ;
  schema:roleName          LITERAL
                        OR [<http://data.europa.eu/esco/>~]              + ;
  m4s:companyShares        LITERAL                                       ? ;
  m4s:isAccumulation       ["true"^^schema:Boolean "false"^^schema:Boolean] ;
```

Figure 16 Extract from the Map4Scrutiny Shapes in ShExC and DCTAP - Roles Reified Statement

The last section maps the TRUE and FALSE pairings into one of the four chars for cardinality in ShExC: empty is the default and means it is mandatory, not repeatable; ‘?’ is not mandatory, not repeatable; ‘+’ is mandatory, repeatable; ‘*’ is not mandatory, repeatable. Another char for the cardinality could have been added before the “OR” to specify different cardinalities for the literal and the vocabulary, something that is not yet possible in the DCTAP tabular format.

Going back to the middle section for a final note. The value shapes are not represented due to a validation constraint explained in 4.3.2 Validation. To specify them the syntax would be “IRI @:Organization” indicating that the IRI that serves as value must conform to the shape named :Organization.

4.3 Transformation, Validation and Publishing

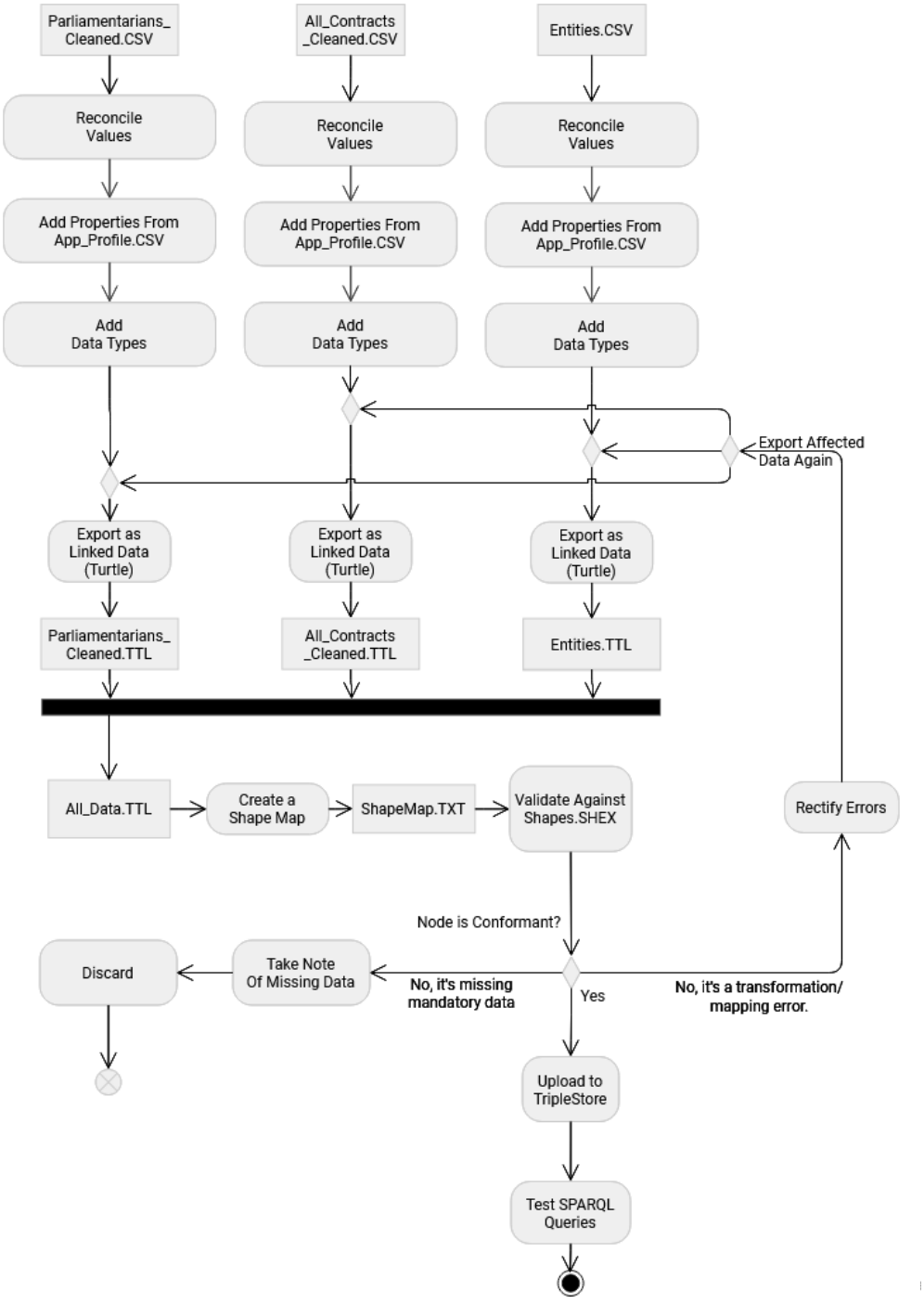


Figure 17 Transform Validate and Publish

OpenRefine was used to map the data into what was defined in the application profile Table 21. Figure 17 represents this process. The three tabular files with data about the Parliamentarians, Organizations and Processes all go through the same steps until they are exported in a Turtle format.

4.3.1 Transformation

The first step is transforming the values that are meant to be a vocabulary into a link. This is done with the OpenRefine reconciliation tool. Essentially what this tool does is that it uploads vocabularies and then, for each column, the user selects the vocabulary in which OpenRefine should look for the values. OpenRefine reconciles to values of the column selected with the data it has from the vocabulary selected and returns possible matches with a degree of certainty. Finally, the user reviews these matches and either accepts, changes or simply refuses them. The last option leaves a string value.

In Table 24 there is a list of all properties that can have a controlled vocabulary as a value, the total of values for that property, and how many were not linked to the vocabulary. The largest number of "not linked" values are in the ESCO vocabulary for occupations and the "Kind" of contract. The first was not matched when an equivalent to the described occupation was not found or when the described occupation did not have enough detail to be linked to the vocabulary. For instance: "Retired" or "Manager". The second term is too broad to find a suitable link in the ESCO vocabulary.

The "Kind" of the contract is also a controlled vocabulary. Nevertheless, it allows for the value "Other kind" followed by a description. These cases are singular and therefore not included in the Contract Kinds vocabulary.

Table 24 Reconciling Controlled Vocabularies

Subject	Predicate	Value Vocabulary	Total Values	Not Linked
	Gender	Schema Gender	224	
	Property Regimen	Property Regimen	119	
	Has Occupation	ESCO	234	44
	Commissions	Parliamentary Commissions	972	
	Constituency	Getty TGN	224	2
Contracts	Procedure Type	*Tipo Procedimento*	131 667	
	Kind	*Tipo Contrato*	133 067	353
	Main Object	CPV	134 322	
	Location	Getty TGN	146 679	1
Entities	Location	Getty TGN	354	

Although it is not included in the table above because it is not a vocabulary, the OpenRefine reconciling tool was also applied to the names of the Parliamentarians to retrieve their profiles on Wikidata. Only one

parliamentarian was not found, the others were appended to the data with the connection "schema:sameAs". This connection makes it possible to get any extra info that Wikidata has on each parliamentarian when using SPARQL to query the final dataset.

Having the values all set, the next step is mapping the columns to the properties, predicates, defined in the application profile. Overall, each column is a property except for properties that allow both URI and Literal Values. Those cases have two columns where one is mapped as a URI and the other as a Literal. This happens mainly with controlled vocabularies such as "m4s:constituency". The value can either be a member of the Getty vocabulary for a Portuguese district or the string statement "Outside of Europe".

An example of this process is presented below in Figure 18 and above in Figure 13. In front of the property name, is the indication of the column and an indication of the datatype. The names of the columns resemble those of the properties for accelerating the mapping process.

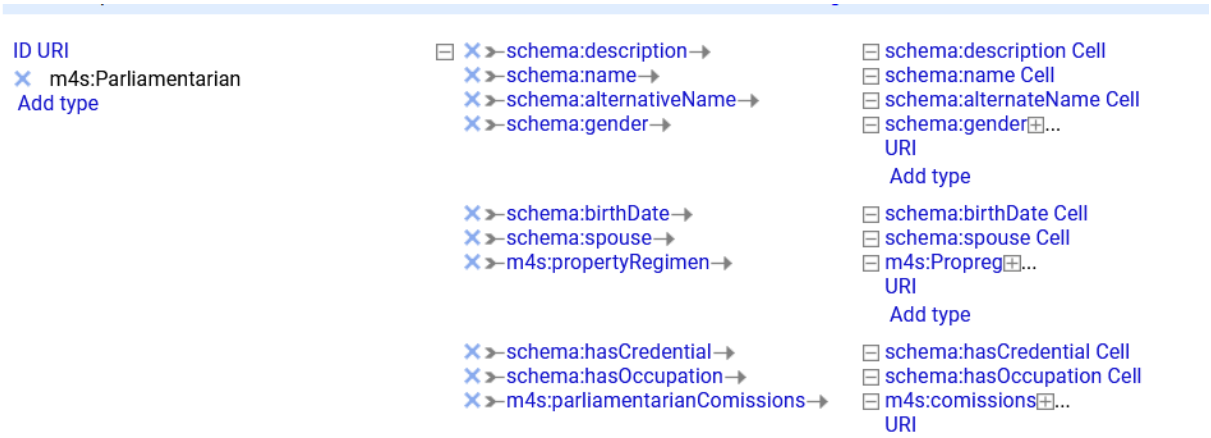


Figure 18 Map4Scrutiny OpenRefine Print – Mapping from Columns to RDF – Parliamentarians

After this step, exporting the data is only one click away. The resulting files were then merged into one file with all the triples of the project. Below, Figure 19 shows one subject, a parliamentarian, and is a sample of the Turtle syntax generated from the map in Figure 18. After the "^^" are the datatypes chosen that were mapped in OpenRefine but could not be seen in Figure 18. This sample has all the properties and is a member of several organizations but that is not always the case. The blurred areas correspond to personal data.

```
<https://example.org/map4scrutiny/entity/[redacted]> a m4s:Parliamentarian;
  dbo:parliamentaryGroup "[redacted]";
  m4s:constituency <http://vocab.getty.edu/tgn/7003826>;
  m4s:parliamentaryComissions <https://example.org/map4scrutiny/value#CAPMADPL>, <https://example.org/map4scrutiny/value#CS>;
  m4s:parliamentaryTerm "XIV";
  m4s:propertyRegimen <https://example.org/map4scrutiny/value#ComAdquiridos>;
  schema:alternateName "[redacted]";
  schema:birthDate "1978-08-07"^^schema:Date;
  schema:comment "[redacted]";
  schema:description "Deputado à Assembleia da República";
  schema:gender schema:Male;
  schema:hasCredential "Licenciatura";
  schema:hasOccupation <http://data.europa.eu/esco/occupation/a580e79a-b752-49c1-b033-b5ab2b34bfba>;
  schema:memberOf <https://example.org/map4scrutiny/entity/O127106>, <https://example.org/map4scrutiny/entity/O198551>,
    <https://example.org/map4scrutiny/entity/O52511>, <https://example.org/map4scrutiny/entity/O800>,
    <https://example.org/map4scrutiny/entity/OL25101>, <https://example.org/map4scrutiny/entity/OL25199>,
    <https://example.org/map4scrutiny/entity/OL25396>;
  schema:name "[redacted]";
  schema:sameAs <https://www.wikidata.org/wiki/[redacted]>;
  schema:spouse "[redacted]";
  schema:url <http://www.parlamento.pt/DeputadoGP/Paginas/Biografia.aspx?BID=[redacted]> .
```

Figure 19 Map4Scrutiny Data Print From Turtle File - One Parliamentarian

4.3.2 Validation

Validation is done essentially by parsing the data generated in OpenRefine and comparing it to the previously defined shapes in ShExC. For this process three files are needed, a Turtle file with the description of all data, a ShEx file with the schema of the data, and a Shape Map. A Shape Map is a simple text file matching each node to its shape, thus informing the validator that Subject A should pass as a parliamentarian while Subject B should pass as an entity. [47] The syntax is very simple, and an example is displayed in Figure 20 Sample of Shape Map for ShEx validation.

```
<Subject_A_URI>@<Shape_Parliamentarian_URI>,
<Subject_B_URI>@<Shape_Organization_URI>,
<Subject_C_URI>@<Shape_Contract_URI>,
...
```

Figure 20 Sample of Shape Map for ShEx validation

The actual validation was run on local implementation of ShEx-java[22] made available on GitHub and it generated an output text file indicating whether the shapes are conformant or not. This file is available for consultation in the support files[23].

The Reified Roles presented a challenge for validation. When parsing the Turtle file with the data, the implementation of ShEx-Java mentioned above gives a new ID to the blank nodes. Let's consider the sample data in Figure 21. The validator starts by parsing the Turtle and ShEx files. When it finds a blank node in the TTL file it gives it a random ID, so <_:node123> could after parsing be <_:nodeA>. The

[22] https://github.com/iovka/shex-java
[23] https://doi.org/10.34622/datarepositorium/KWFSU2

challenge of this characteristic is that when the validator reads the shape map below, _:node123 does not exist anymore and it is not validated.

```
TTL
    _:node123 rdf:type rdf:Statement.

ShEx
    :roles BNODE {
    rdf:type rdf:Statement;
    }

Shape Map
    _:node123@:roles
```

Figure 21 Sample Validation of Blank Nodes

Since the reified statements are blank nodes that do not exist as a value for any property, the solution to validate them against the defined shape was using an online validator[24] that does not give blank nodes new IDs when parsing them and accepts blank node IDs in its shape map. The online validator accepts a shape map such as the one in Figure 21 and outputs a list of conformant and non-conformant shapes. Above,

Figure 16 shows the last version of the shape for the reified roles. The reason why the shape identifier for the Parliamentarians and Organizations had to be removed, was because this data was not present in the online validator. The dataset was too large to be validated entirely there. Nevertheless, all organizations and politicians were validated in the local implementation of ShEx-Java.

The resulting file from the role's validation returned 10 non-conformant roles that were not uploaded because the original data has neither an entity where the role is played nor a role name. Not having an entity means there is no Object which is a part of the reification vocabulary and therefore is mandatory. Not having a Role Name renders the reified statement useless since its purpose is to describe the role.

As may be seen in Figure 17 after the validation two different things can happen to non-conformant data. Mapping and transformation errors refer to mistakes in the previous processes that caused a problem. Two specific samples are the following: A typo when uploading the schema ontology. In the data, it was "HTTP" and in the shapes "HTTPS" this caused several nonconformant nodes that were easily fixed; Some whitespaces were left in the Contracts URI's, this was another error that was easily fixed by removing the spaces and exporting the data again.

After the mapping errors were corrected the data regarding the politicians and the entities all passed the validation process. The data regarding the contracts has 135 nonconformant values. 11 of these are

24 https://shex.io/webapps/shex.js/doc/shex-simple.html

nonconformant because there is no link for the supplier. The supplier is simply a string and has no unique fiscal identifier (NIF), for this reason, it was not possible to include it in the automatic search for entity's links and it does not have one. The remainder 124 are contracts that do not have a common procurement vocabulary (CPV) number. Both situations could have been made valid by making the original shapes even more open, however, having a link to the contractor and supplier is the minimum one could ask for when describing a contract as LOD. Since these are public, having a CPV should also always be required. So, these nodes are not valid against the predicted shape of the data and were not uploaded into the Triplestore.

4.3.3 Publication

The second to last step mentioned above is uploading the data to a Triple Store. Following the recommendations in the literature, this project was uploaded to a local version of OpenLink Virtuoso [50], [4], [37] to test SPARQL queries and verify if the subjects were properly linked to each other.

Table 25 The Final Triples Database in Numbers

Types	Subjects/Entities	Triples
Parliamentarians	224	5 850
Organizations	25 519	101 866
Contracts	131 531	1 898 482
Roles	1650	14 942
Contract Modification and Extinction	52285	143 958
Total	211 209	2 165 098

After discarding the non-conformant data the uploaded triples describe the entities, parliamentarians, and contracts with a total of 2 165 106 triples. Table 25 considers only the data about these but that is not all that was uploaded to the database. To make querying the data in a local implementation a friendlier experience all the controlled vocabularies for values and vocabularies for properties used to describe the data are also uploaded. The summary of their dimension is represented in Table 26. The TGN Location Taxonomy is not fully available. Due to the large volume of the Taxonomy, only the used locations are uploaded in the local implementation. A folder with all the data needed to upload the same data in any local machine is available in the support files.[25]

[25] https://doi.org/10.34622/datarepositorium/K1DQIT

Table *26 Auxiliary Vocabularies*

Vocabulary Name	Subjects	Triples
Map4Scrutiny New Properties	58	255
Map4Scrutiny New Values	84	466
Getty Taxonomy Geographical Locations	30341	357697
ESCO – European Occupations Taxonomy	2950	14076
Common Procurement Vocabulary	10420	298583

The SPARQL endpoint offered by OpenLink Virtuoso allowed running a set of queries both to verify if the data is linked properly and explore the potential of the added vocabularies. With all the vocabularies linked it is possible to query for information that was not available before. For instance, it is possible to have a query retrieving the location of an Organization, the ISO code of that location, what broader location it belongs to and any other property that is described by the Geographical Places Taxonomy. Another example is the common Procurement Vocabulary where it is possible to navigate to narrower or broader connections of a given code.

Figure 22 shows a long query and its return with only six selected values. This query was formulated to showcase several possibilities of the data, not to answer a specific question. The query can de be deconstructed as follows: Any parliamentarian that fits the criteria is represented by the wildcard ?s. The parliamentarian should have a name and a memberOf property. The Organization he is a member of, represented by ?o needs to have a legal name, a NIF and a location. The location, ?heads, cannot be a string because it needs to have a label. Then to get further information on the member of relation, the ?b represents de reified statement that has the same ?s as a subject, the ?o as an object, has also a role name, and the description is "Acionista".

So far, the query is pulling every parliamentarian, that is a shareholder in an organization and that has all the properties listed above. From the reified statement the query also pulls the value for the held shares. Then, the subject changes to ?c, which stands for contracts. The query is now pulling subjects that have ?o as a supplier. These also need to have an actual price, a CPV, a description, and an IRI location that has a label. Both the location and the CPV vocabularies are external and available in several languages. CPV was left as an IRI in these samples, but labels for the location of both the organization's headquarters and the contract execution were pulled and are in the "SPARQL | HTML 5" table.

The labels for the locations have a language filter because otherwise random languages would be presented.

For this sample, only six values are selected to be shown, but any instance of an "?" could be selected and printed. The names, NIFs, and roles were left out of the selection on purpose to avoid singling out personal data.

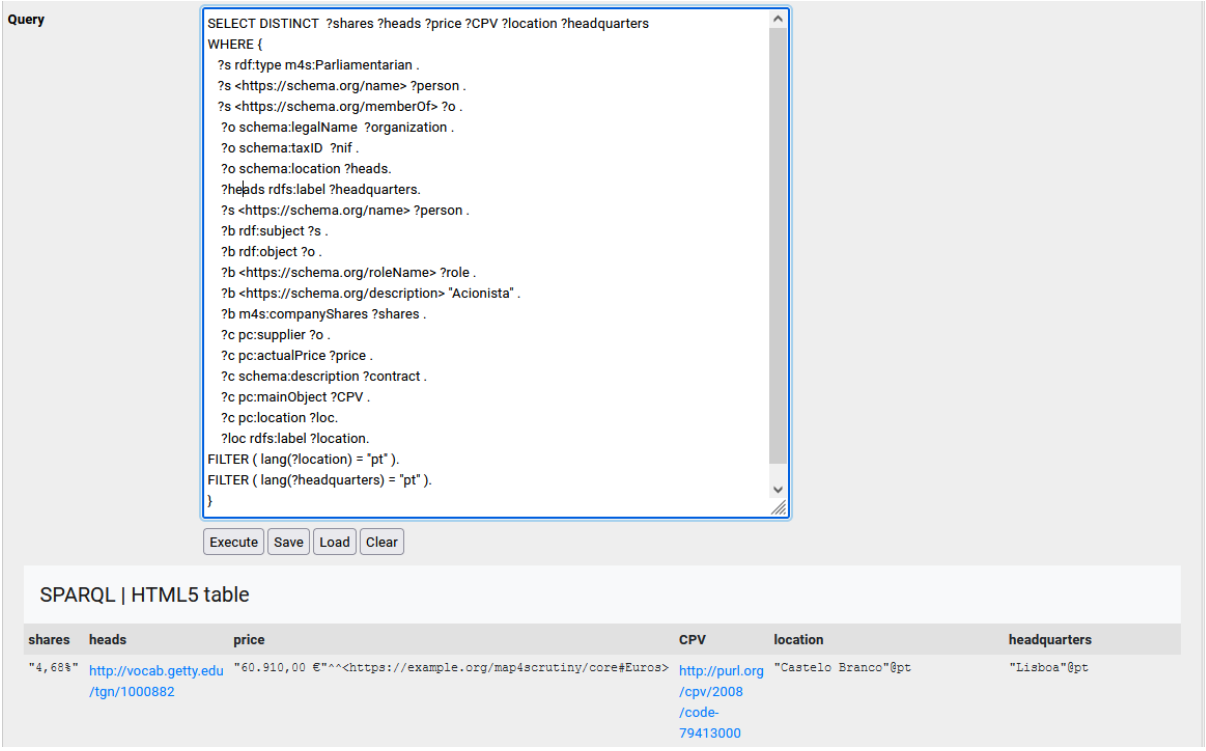


Figure 22 Map4Scrutiny Print OpenLink Virtuoso - Sample SPARQL Query

Other samples of queries are available in Appendix D Sample SPARQL Queries. These samples are meant to showcase the added exploration possibilities that the transformation brought to the original data, they are "link" oriented instead of results-oriented. They can be run by anyone that uploads the bulk data in a Triplestore. Being able to query the data across different vocabularies is the final step of the implementation process.

5 CONCLUSIONS

This chapter is an overview and direct comparison between what was initially planned and what was achieved. There is also a focus on the development experience, the challenges encountered, and the value it added to the design science research's knowledge base and the application domain.

5.1 Context

In the Introduction one main objective and two sub-sequent goals were set: transform open data on Portuguese parliamentarians interest registers and public procurement from their current level (one or two) to level five [2], or LOD; Publishing the resulting data online via a SPARQL endpoint; Make the process transparent so that any third party could replicate this solution.

The first objective is technical. The datasets used are part of the Table 6 Priority datasets [34] and key datasets [18] defined by both the Open Data Charter and the G20. The interest registers were available in HTML and the contracts in CSV, so they were considered one star, and three stars, respectively [2]. In the final dataset, the data from both sources is connected to each other and to external data that it had no contact with before, such as the CPV, locations, and occupations vocabularies. Moreover, it provides further context on the original with new vocabularies created within the scope of the solution. It is, therefore, by definition, five stars OD [2]. The final solution is currently missing one characteristic of LOD which is that the dataset is not yet published on the web.

This step was not implemented because the dataset, ontology and vocabularies are not published online RDF files. They are only uploaded in a local instance of the Virtuoso Triplestore and available in bulk for anyone else to upload them to their local Triplestore or to publish them online. [26]

The data was meant to be published in a university's server with a working SPARQL endpoint. This publication was not possible due to technical constraints not related to the work developed. Nevertheless, the data is available in bulk for download and ready to be uploaded to a local or online instance of a Triplestore by any party.

For the same reason, the implementation process employed all the features in Table 9 Linked Data Checklist from "How to use Linked Data"., except for not yet having the data registered in Linked Data Catalogues.

[26] https://doi.org/10.34622/datarepositorium/K1DQIT

As for the third objective, the whole process has been described and all the data and code are available for both consultation and re-use.

By fulfilling the objectives, the proposed solution is also a suitable solution for the stated research problem: The data provided by the government on Parliamentarians Interest Registers and Public Procurement is neither machine-readable nor interoperable by semantic web standards. The transformed dataset is both interoperable and connected to external data, and machine-readable.

In summary, the first and third objectives were fully accomplished, the second partially accomplished in the sense that the steps that could be taken within the scope of this work to fulfil that objective are completed.

The remainder of chapter five explores how the outputs that contributed to designing and building the proposed solution contribute back to the domain application and the knowledge base, followed by a set of recommendations for the data sources. At last, based on the questions raised by the research and implementation stages, future work is proposed to further expand bot the knowledge base and application domain.

5.2 Contributes

The research and outputs that fulfilled the objectives above also constitute contributions both to the knowledge base and the application domain.

5.2.1 Knowledge Base

This chapter makes an overview of the contributions the work developed has for the knowledge base identified in Figure 3: Drechsler, Hevner (Ed.) 2016 - A three-cycle view of design science research [6].

The relevance of this work affects mainly the LOD and Semantic web research fields by providing a detailed description of an implementation that enables replication independent of the data sources and themes. Every step of the implementation is described, the transformed dataset is open to the public and so are the intermediate code and files needed to achieve the final dataset.

For the grounding of the work, every software, step, and decision was based on literature and prior art. The goal was to create a sustained solution that follows LOD and Semantic Web guidelines and to document a full LOD transformation process.

The work developed contributions back to the foundations on semantic web with new Linked Data vocabulary of properties, classes and a datatype properly connected to existing vocabularies, controlled

vocabularies of values for the suitable topics, annotations with translation in re-used properties and classes. Still in the semantic web field fits the final Dataset in Turtle, the prior-art research, the overview on reification methods, an LD rendition of the European vocabulary for occupations in Portuguese, the tabular application profile for the data, the shapes of the data expressed in ShExC, a description of the validation process together with the files used and needed to replicate the said process. In the field of transparency and OD fit the python scrappers used to retrieve data both from Parlamento.pt and Base.Gov, the added value to data when the quality is increased, the application profile, and the research on prior-art.

5.2.2 Application domain

The interest of the public in the originally selected datasets was reinforced by the results of the Global Corruption Barometer 2021 and the reactions to a recent blackout of the Base.gov.pt portal. The barometer shows that 90% of the Portuguese population believes there is corruption in the government and 27% stated that most parliamentarians are corrupt. This places parliamentarians as the perceptibly more corrupt government seats. The drop is significant when it comes to members of government, these are only believed to be corrupt by 16% of the population. [66]

As for the Base.gov.pt portal blackout, between the first and the fourth of October 2021, the portal Base.gov.pt was unavailable due to GDPR infringements related to the disclosure of personal information in the procedure pieces. On the fourth, the portal was put back online without access to any procedure features. [67] This episode reminds the public of the frailty of the access to OD that is currently available. The data in the contracts is often incomplete, and without the procedure pieces, that are still not available, access to public procurement is limited. [67] As a result of this episode the property “schema:URL” of the contract's description that linked to the procedure pieces was also removed from the bulk data made available as an output of this project.

The research conducted in this dissertation and the proposed solution both show that the current data made available on parliamentarians could be much richer and valuable without having to disclose more information, only by disclosing it with context, quality, and in a machine-readable format. The same could be said for the contracts, these could take advantage of ranking as five stars, having better quality control and not remaining so dependent on procedure pieces and other unstructured data documents.

In the areas of OD and Transparency, the solution shows how information that is already available to the public could have added value if on top of availability there is also work on quality and linking. Choosing data that, according to surveys, is likely to be of interest to the public, despite the low quality it is made

available in, presents challenges and ads work but also creates a higher step between what was scraped or retrieved and what is uploaded to the Triplestore in the end.

5.2.2.5 Recommendations

One of the largest challenges when developing the proposed solution was the quality of the data, but despite the difficulties faced, the cleaning process generated a set of observations on the data. These were then compiled with good practices on OD and resulted in a set of recommendations for both Parlamento.pt and Base.gov.pt presented below, preceded by a review on the data quality issues that propelled the recommendations.

In Parlamento.pt that data appears to not be reviewed at all and there is no coherence between the registers. The values for certain properties were not appropriate and many values resulted in entities or positions that could not be disambiguated. An example of information that may be misleading is: There are values with the comment "no salary" under "roles" in social and other working positions. This comment was added in front of the role name. Not receiving a salary is not part of a job name, and since this information is not asked for by the form, it is wrong to assume that every other role that does not specify being salaried is therefore salaried.

As for the contracts, the most noteworthy gaps in information related to the existence of contracts without a CPV or without a supplier being identified with at least a NIF. The database for Base.gov.pt is a lot larger than the interest registers from Parlamento.pt and, for this reason, the challenges in maintaining data quality are in no way comparable. Nevertheless, some basic requirements should be met. Some examples found here that caused ambiguity problems are: Several profiles for the same entity, the same entity existing with different NIFS, and several variations of "Nothing changed" strings written as values for price and procedure changes.

Combining the experience in cleaning the data with the good practices retrieved from the literature resulted in a further contribution to the knowledge base in the form of a set of recommendations both Base.gov.pt and Parlamento.pt could implement that would be an incentive to data re-use and analysis by third parties.

In Parlamento.pt both the registers from the parliamentarians and the members of the government should be made available in a bulk format as they do with biographical information. As for the quality of the data, there should be a set of rules, explanations, and standards to avoid misplaced data, the registers should all be checked for quality before being published and some controlled vocabularies could be added such

as the EU taxonomies for qualifications and occupations. It would also be ideal that all mentioned organizations are identified uniquely, this could be done with a NIF.

In Base.gov.pt the most important recommendation would be that all contracts should have at least the name and NIF of the involved parties and a CPV to be published. For the latter, if there are any specific cases where a CPV is not appliable, this should have a justification comment. As for the remainder of the inconsistencies found, they relate closer to overall data quality. Nevertheless, they affect data analysis, for instance, the price and date modification feature with variations of the string "nothing changed" create noise and create an obstacle to simple statistics such as querying "how many contracts had a modification" by counting not empty strings.

Experience suggests that to have good quality LOD on public procurement and interest registers, the most crucial factor, is the quality of the data available. It is necessary for OD portals to improve their standards for quality and to display data with context that may be analysed by any individual without the risk of making assumptions.

5.3 Future Work

Future work is also divided into two sections. Refinements that were not implemented because the time it would take was found to outweigh the advantages, or due to time constraints. Despite this, the actions needed to implement these additions are clear and described below. The second section relates to more challenging work and dissemination. This work is now in a laboratory stage, and a further process would also be to test its impact with the identified stakeholders and evaluate the relevance of future improvements.

Following the review of the objectives, the first task on the future work would be to upload the data to a server that maintains an online SPARQL endpoint for the public to explore.

Following that, and being this a proposed solution, not all challenges were solved in a manner that can be simply mirrored into a broader implementation. Below are the features identified during the research and implementation processes that are suitable to constitute future work.

The first suggestions are related to maintaining the dataset. A lot of manual work went into creating the proposed solution and it is not viable to keep the solution up to date with the same process.

Automating the data sourcing process - The Scraping process could be easily improved and automated into a scrapper that runs regularly and retrieves new data when there is any.

Automating the data cleaning process – This was the task that had the most manual working hours, and currently, keeping the data updated without manual cleaning is not possible. The transformations made would need to be further refined to create an extraction, transform, and load (ETL) pipeline.

Improve Value vocabularies - The value vocabularies related to specific legal matters are only available in Portuguese and named in Portuguese because an official direct translation was not found. These should be officially translated and enriched with legal citations. Another future work task is to develop a more stable and complete version of the ESCO SKOS vocabulary. The one available here does not include every language nor every broader property. The description that ESCO makes available in bulk has data that allows for a more complete solution.

Broadening the Data Scope - The application profile for the contracts could be improved with more information and closest alignment with the Public Procurement and Public Contracts ontologies. Currently, there are no contracts described with more than one contracting or supplier entities, but the application profile supports these being brought into scope. Furthermore, it would be interesting to see related organizations being linked, for instance, a University and their Social Services. These are two different entities with different NIFs that work closely together, and these connections should be mapped. Another feature that could be added is a schema:sameAs for entities, just like it was done for the parliamentarians. There is a high chance that at least the municipalities have a profile in Wikidata with more information than what is available in this project.

The State of the art also left open more questions on the use of LOD as a corruption-fighting tool than this work could close. Currently, it is neither possible to state whether LOD can have an impact on corruption fighting or not.

Overall, the use of LOD as a possible corruption-fighting tool via transparency improvement shows potential but further work and projects need to be developed to draw further conclusions on the value of such implementations.

The challenge with corruption is that it is often measured by the perception of corruption by the public. This is one of the reasons why understanding if good OD policies (that include LOD projects) have an impact on the perception of corruption, is a field that needs further studying because there is a very heavy social component to it that requires testing and experimenting with the public.

5.4 Closing Note

In the end, the work developed during this dissertation proposes a solution that constitutes an improvement on the original proposals. The proposed solution has a large room for improvement but above all, it shows that there is a future for LOD in increasing transparency. The final dataset feeds from isolated interest registers from parliamentarians, and public contracts that exist as static documents and, by creating a unique identifier for organizations, links them to the organizations they have in common. This simple idea that is then worked on with further context and details completely changes the potential for re-use and querying that the original data had. This feature describes the value that five-star OP has in comparison to any of the previous levels. To have truly open data and achieve transparency with it, the discussion cannot rest only on the number of topics made available and if these topics are of interest to the public, but also on the quality and format of the data.

6 APPENDIX

Appendix A Map of Data available in Parlamento.pt and Base.gov

Grupo de dados	Títulos	Propriedades	Exemplo preenchimento	Incluir informação
Perfil		Nome completo	Adão José Fonseca Silva	
		Data de Nascimento	01/10/1957	
		Habilitações literárias	Licenciatura em Línguas e Literaturas Modernas - Estudos Portugueses e Franceses	
		Profissão	Professor	
		Cargos que desempenha	Deputado na XI Legislatura;	
		Cargos exercidos	Secretário de Estado Adjunto do Ministro do Saúde.::Deputado na V, VI, VIII, IX e X Legislaturas::Presidente da Assembleia Municipal-Macedo de Cavaleiros::Professor efetivo no ensino secundário	Redundante e para os que aconteceram há mais de três anos não há mais informação
		Comissões Parlamentares a que pertence	Comissão de Defesa Nacional	
		Grupo Parlamentar / Partido	PSD	
		Círculo eleitoral	Bragança	
		Legislatura	V::VI::VIII::IX::X::XI::XII::XIII::XIV	
Presenças às Reuniões Plenárias		Data	20021-02-25	Fora do Scope deste projeto
		Número	47	Fora do Scope deste projeto
		Tipo	Ordinária	Fora do Scope deste projeto
		Presença/Falta	Presença	Fora do Scope deste projeto
Atividade do Deputado	Atividade	Iniciativas apresentadas		Fora do Scope deste projeto
		Tipo	Apreciação Parlamentar	Fora do Scope deste projeto
		Número	36/XIV	Fora do Scope deste projeto
		Sessão	2	Fora do Scope deste projeto
		Título	Decreto-lei n.º 102-D/2020, de 10 de dezembro, que aprova o regime geral da gestão de resíduos, o regime jurídico da deposição de resíduos em aterro e altera o regime da gestão de fluxos específicos de resíduos, transpondo as diretivas (UE) 2018/849, 2018/850, 2018/851 e 2018/852	Fora do Scope deste projeto

Atividade	Requerimentos apresentados		Fora do Scope deste projeto
	Número	185/AC/XIV/2	Fora do Scope deste projeto
	Data	18/02/2021	Fora do Scope deste projeto
	Título	Impactos de Exploração Mineira junto da Fronteira Portuguesa	Fora do Scope deste projeto
	Atividade	Perguntas apresentadas	Fora do Scope deste projeto
	Número	1403/XIV/2	Fora do Scope deste projeto
	Data	18/02/2021	Fora do Scope deste projeto
	Título	Impactos de Exploração Mineira junto da Fronteira Portuguesa	Fora do Scope deste projeto
Atividade	Comissões a que pertence / pertenceu		Fora do Scope deste projeto
	Comissão	Comissão de Negócios Estrangeiros e Comunidades Portuguesas [Suplente]	Fora do Scope deste projeto
	Atividade	Intervenções	Fora do Scope deste projeto
	Data da reunião	27/10/2020	Fora do Scope deste projeto
	Legislatura	XIV	Fora do Scope deste projeto
	Sessão	2	Fora do Scope deste projeto
	Tipo	Pedido de esclarecimento	Fora do Scope deste projeto
	Sumário	Lei das Grandes Opções para 2021-2023; Aprova o Orçamento do Estado para 2021	Fora do Scope deste projeto
Atividade	Atividades parlamentares		Fora do Scope deste projeto
	Tipo	Voto	Fora do Scope deste projeto
	Número	475	Fora do Scope deste projeto
	Legislatura	XIV	Fora do Scope deste projeto
	Sessão	2	Fora do Scope deste projeto
	Data de entrada	24/02/2021	Fora do Scope deste projeto
	Data do debate	25/09/2020	Fora do Scope deste projeto
Atividade	Delegações Eventuais - Reuniões em que participou		Fora do Scope deste projeto
	Legislatura	XIV	Fora do Scope deste projeto
	Sessão	1	Fora do Scope deste projeto
	Local	Roma	Fora do Scope deste projeto

		Inicio	11/11/2019	Fora do Scope deste projeto
		Fim	14/11/2019	Fora do Scope deste projeto
		Delegação	Visita de Estado à República Italiana	Fora do Scope deste projeto
	Atividade	Audições		Fora do Scope deste projeto
		Número	27-CDN-XIV	Fora do Scope deste projeto
		Data	02/12/2020	Fora do Scope deste projeto
		Comissão	Comissão de Defesa Nacional	Fora do Scope deste projeto
		Assunto	Audição da Comissão de Trabalhadores da Arsenal do Alfeite, SA, a requerimento dos Grupos Parlamentares do PSD BE e PCP, a fim de esclarecer este Parlamento sobre a situação da empresa	Fora do Scope deste projeto
		Entidades	Presidente Rui Ferreirinho; António Pereira; Tibério Rodrigues; Hugo Pereira	Fora do Scope deste projeto
Registo de Interesses		Nome completo	ADÃO JOSÉ FONSECA SILVA	Redundante Linha - 2
		Atividade principal	Deputado à Assembleia da República	Redundante Linha - 6
		Estado civil	Casado(a)	
		Nome completo do cônjuge	Ana Maria Afonso Silva	
		Regime de bens	Comunhão geral	
		Cargo/função	Deputado à Assembleia da República	Redundante Linha - 6
		Alteração:	17/09/2020	
	Cargos/funções/ atividades	Últimos três anos		
		Cargo/função/atividade	Deputado à Assembleia da República	
		Entidade	Assembleia da República	
		Início	23/10/2015	
		Fim	24/10/2019	
		Cargo/função/atividade	Presidente do Grupo Parlamentar do PSD	
		Entidade	Assembleia da República	
		Início	17/09/2020	
		Fim		
	Cargos/funções /atividades	Acumulação com cargo político/alto cargo público		
		Cargo/função/atividade	Membro da Assembleia Municipal da Trofa	
		Entidade	Assembleia Municipal da Trofa	
		Início	22/10/2017	
		Fim		
	Cargos/funções/ atividades	Até três anos após cessação de funções		não aplicável pois estamos só a lidar com os atuais

	Cargos sociais	Últimos três anos	
		Cargo	Gerente (não remunerado)
		Entidade	Clínica Médica Dentária Glória Ferreira, Lda.
		Natureza e área de serviço	Medicina Dentária e Estomatologia; Saúde Humana
		Local da sede	Rua D. Afonso Henriques, 125 - 1º andar - sala N - 4435-005 Rio Tinto
	Cargos sociais	Acumulação com cargo político/alto cargo público	
		Cargo	Presidente da Mesa da Assembleia da Delegação Distrital do Porto (não remunerado)
		Entidade	ANAFRE - Associação Nacional de Freguesias
		Natureza e área de serviço	Associativa
		Local da sede	Porto
	Cargos sociais	Até três anos após cessação de funções	não aplicável pois estamos só a lidar com os atuais
	V - Apoios ou benefícios		Senhas de Presença Comissão de Ética para a Investigação Clínica: 160€
	VI - Serviços prestados		Comentadora política TVI Fevereiro 2018 a Julho 2018 Comentadora política SIC Novembro 2018 a setembro 2019
	VII - Sociedades		será sempre uma Spring porque é preenchido em forma de comentário
		Entidade	Clínica Médica Dentária Glória Ferreira, Lda.
		Área da atividade	Medicina Dentária e Estomatologia; Saúde Humana
		Local da sede	Rua D. Afonso Henriques, 125 - 1º andar - sala N - 4435-005 Rio Tinto
		Participação social	74%
	VIII - Outras situações		As atividades acima referidas em ONG, como Investigadora e como Presidente das Mulheres Socialistas não são remuneradas.
	IX - Declaração sobre exclusividade		Exclusividade/ Não exclusividade

Grupo de dados	Propriedades	Exemplo Preenchimento	Incluir Informação
Contratos	Data da publicação	24/02/2021	
	Tipos de contrato	Aquisição de serviços	
	Nº do acordo quadro	Não aplicável.	
	Descrição do acordo quadro	Não aplicável.	
	Tipo de procedimento	Ajuste Direto Regime Geral	

	Descrição	Prestação de serviço para a criação de campanha de Marketing Digital	não incluída na exportação CSV e Redundante com **A20**
	Fundamentação	Artigo 20.º, n.º 1, alínea d) do Código dos Contratos Públicos	
	Fundamentação para recurso ao ajuste direto (se aplicável)	ausência de recursos próprios	não incluída na exportação CSV
	Entidades adjudicantes	Universidade Nova de Lisboa (501559094)	
	Entidades adjudicatárias	Mosca Azul Publicidade, Lda. (510855679)	
	Objeto do contrato	Prestação de serviço para a criação de campanha de Marketing Digital para promoção e divulgação do novo programa de bolsas dos SASNOVA.	
	Procedimento centralizado	-	não incluída na exportação CSV
	CPVs	79342000-3, Serviços de marketing	
	Data do contrato	10/02/2021	
	Preço contratual	19.860,00 €	
	Prazo de execução	182 dias	
	Local de execução	Portugal, Lisboa, Lisboa	
	Entidades concorrentes	-	não incluída na exportação CSV
	Anúncio	-	não incluída na exportação CSV
	Peças do procedimento	-	não incluída na exportação CSV
	Modificações contratuais	-	não incluída na exportação CSV
	Documentos	Contrato_Mosca_rasurado.pdf	é uma cópia do contrato não normalizada
	Observações	-	não incluída na exportação CSV
	Critérios ambientais	-	não incluída na exportação CSV
	Justificação para não redução a escrito do contrato	-	não incluída na exportação CSV
	Causa da extinção do contrato	-	
	Data do fecho do contrato	-	
	Preço total efetivo	-	
	Causas das alterações ao prazo	-	
	Causas das alterações ao preço		
Contratos CSV Extra	Estado		
	N.º registo do Acordo Quadro	Não aplicável.	
	Descrição do Acordo Quadro	Não aplicável.	

	Ligação para Peças do Procedimento		
	Lista de Fornecedores (cocontratantes)	FALSE	
Entidade	NIF	510855679	redundante **A18 e A19**
	Descrição	Mosca Azul Publicidade, Lda.	redundante **A18 e A20**
	Localização da sede	Portugal	Redundante info parlamento
	Nº de contratos como adjudicante	0	count function
	Total gasto	0,00 €	sum function
	Contratos como adjudicante	-	count function
	Nº de contratos como adjudicatária	5	count function
	Total ganho	194.185,00 €	sum function
	Contratos como adjudicatária	Lista dos contratos	link para toda a info **A9 - A39**

Appendix B CSV headers with selected data for the Application

Grupo de dados	Propriedades	Tipo de Dados	Exemplo preenchimento	
Perfil	Nome Completo	string (1)	Adão José Fonseca Silva	Sobre o Deputado como descritivo
	Data de nascimento	data (1)	01/10/1957	
	Habilitações literárias	string (1,n)	Licenciatura em Línguas e Literaturas Modernas - Estudos Portugueses e Franceses	
	Profissão	string (1,n)	Professor	
	Comissões Parlamentares a que pertence	String (0,1,n)	Comissão de Defesa Nacional	
	Grupo Parlamentar / Partido	String [0:1]	PSD	
	Círculo eleitoral	String [1:n)	Bragança	
	Legislatura	string [1,n]	V::VI::VIII:IX::X::XI::XII::XIII::XIV	
Registo de interesses	Estado civil	String (1)	Casado(a)	
	Nome completo do cônjuge	String (0,1)	Ana Maria Afonso Silva	
	Regime de bens	String (0,1)	Comunhão geral	
Perfil	Cargos/funções/atividades	String (1,n)	Deputado à Assembleia da República	Conexão entre Deputado e Entidade
	Acumulação	Boolean(True/False)	FALSE	
	Início	data (1)	17/09/2020	
	Fim	data (0,1)	24/10/2019	
	Tipo Cargo	Cargo Social, Cargo, Função ou Atividade, Sociedade	Cargo Social	
	Participação social	Percentagem (n) (só aplicável quando tipo = Sociedade)	74%	
	Entidade	String (0,1,n)	Assembleia da República	Sobre entidade
	Natureza e área de serviço	String (0,1,n)	Medicina Dentária e Estomatologia; Saúde Humana	
	Local da sede	String (0,1,n)	Rua D. Afonso Henriques, 125 - 1° andar - sala N - 4435-005 Rio Tinto	
	Apoios ou benefícios	será sempre uma *string* porque é preenchido em forma de comentário	Senhas de Presença Comissão de Ética para a Investigação Clínica: 160€ Senhas de Presença Conselho Nacional para a Procriação Medicamente Assistida: 70€	Sobre deputado
	Serviços prestados	será sempre uma *string* porque é preenchido em forma de comentário	Comentadora política TVI Fevereiro 2018 a julho 2018 Comentadora política SIC Novembro 2018 a setembro 2019	
	Outras situações	será sempre uma *string* porque é preenchido em forma de comentário	As atividades acima referidas em ONG, como investigadora e como Presidente das Mulheres Socialistas não são remuneradas.	
	Declaração sobre exclusividade	Boolean: Exclusividade/ Não exclusividade	Exclusividade	

Grupo de dados	Propriedades	Tipo de Dados	Exemplo preenchimento (CSV gerado
CSV Contratos por entidade	Objeto do Contrato	string comentário	Prestação de serviço para a criação de campanha de Marketing Digital para promoção e divulgação do novo programa de bolsas dos SASNOVA.
	Tipo de Procedimento	string (vocabulário na forma de lista d opções	Consulta Prévia
	Tipo(s de Contrato	string (vocabulário na forma de lista d opções	Aquisição de serviços
	CPVs	string (vocabulário na forma de lista d opções	79342000-3, Serviços de marketing
	Entidade(s Adjudicante(s	String com NIF (	Universidade Nova de Lisboa (501559094
	Entidade(s Adjudicatária(s	String com NIF (	Mosca Azul Publicidade, Lda. (510855679
	Entidade(s Adjudicante(s	string (Transformed from F)	Universidade Nova de Lisboa
	Entidade(s Adjudicatária(s	int (Transformed From F)	501559094
	NIF(s Adjudicante(s	string (Transformed from G)	Mosca Azul Publicidade, Lda.
	NIF(s Adjudicatária(s	int(Transformed from G)	510855679
	Preço Contratual	int (normalized)	19.860,00 €
	Data de Publicação	date	24/02/2021
	Data de Celebração do Contrato	date	10/02/2021
	Prazo de Execução (dias)	string	182 dias
	Local de Execução	string Geo (pode ser mais que 1 separado por </br>)	Portugal
	Fundamentação	string (vocabulário baseado na lista)	Artigo 20.°, n.° 1, alínea d do Código dos Contratos Públicos
	Causa de Extinção do Contrato	string (campo normalizado - lista valores	
	Data de Fecho do Contrato	date	
	Preço Total Efetivo	int (normalized)	
	Causas das Alterações ao Prazo	(string Comentário)	
	Causas das Alterações ao Preço	(string Comentário)	
	Estado		
	N.° registo do Acordo Quadro	(é um vocabulário)	Não aplicável.
	Descrição do Acordo Quadro	(é um vocabulário)	Não aplicável.
	Procedimento Centralizado	Boolean (true or false)	FALSE
	Ligação para Peças do Procedimento	URL	
	Lista de Fornecedores (cocontratantes	Bollean (true or false)	FALSE

rdfs:domain m4s:Parliamentarian subclassOf: foaf:Person same as: ocd:Parliamentarian

Attribute ER	Property	Description	Original Domain	Original Range	Allowed values/Datatypes	Cardinality
Description	schema:description	known for...equal for all of them	schema:Thing	schema:Text	"Politico Português"	1
Parliamentarian Name	schema:alternateName	Selected two or three names	schema:Thing	schema:Text	LITERAL	1
Name	schema:name	Full name	schema:Thing	schema:Text	LITERAL	1
Gender	schema:gender	Bio Gender	schema:Person	schema:GenderType or Text	Schema:Gender	1
Birthdate	schema:birthDate	Date of birth dd-mm-yyyy	schema:Person	schema:Date	schema:Date	1
Spouse's Name	schema:spouse	Full name of the partner	schema:Person	schema:Person	LITERAL	0-1
Matrimonial property regimen	m4s:propertyRegimen	Distribution of assets in the relationship	schema:Person	x	skos:Concept	0-1
Education Level	schema:hasCredential	Bachelor, Post doc..	schema:Person	schema:EducationalOccupationalCredential	skos:Concept	0-n
Profissão	schema:hasOccupation	Main job	schema:Person	schema:Occupation	skos:Concept	1
Parliamentary commissions	m4s:parliamentary Comissions	Commissions, workgroups and subcommissions from the Portuguese parliament	schema:Person	x	skos:Concept	0-n
Parliamentary group	dbo:parliamentaryGroup	Mainly same as party (except:Coligações)	dbo:Person	x	schema:Text	0-1
Political Party	schema:memberOf	militant this party	schema:Person	schema:Organization	schema:Organization	1
Parliamentary Term	m4s:ParliamentaryTerm	governance term		x	LITERAL	1-n
Electoral cycle	m4s:constituency	electoral district		x	Skos:Concept	1-n?
Entity	schema:memberOf	entity where a position existed/exists	schema:Person	schema:Organization	schema:Organization	0-n
Support and benefits / Services / Commentary	schema:comment	hand writen note on other types of support or perks, on services and commentaries.	schema:CreativeWork, schema:RsvpAction	schema:Comment	LITERAL	1

metainformation for Parliamentarian> memberOf > Organization type:Role

Jobs/Activities	schema:roleName	the position held in a given entity	Role / Event	Text or URL	LITERAL	0-1
End Date	schema:endDate		Role	Date	Schema:Date	0-0
Start Date	schema:startDate		Role / Event	Date	Schema:Date	0-1
Shares	m4s:companyShares	owned shares in % or €	schema:Person, schema:Organization		LITERAL	1
Accumulation	m4s:isAccumulation	happened at the same time as the current mandate		schema:Boolean	schema:Boolean	1
Work Relation	schema:description	If it is a social position, a job or a society	schema:Thing	schema:Text	LITERAL	

schema:Organization

company name	schema:legalName	name of company	schema:Organization	schema:Text	LITERAL	1
area of activity	schema:hasOfferCatalog	typed area of activity for societies only	schema:Organization	schema:OfferCatalog	LITERAL	0-1
headquarters	schema:location	where it is located	schema:Organization	Place,Text and other	skos:Concept	0-1
NIF	schema:taxID	9 digit unique identifier number	schema:Organizatio, schema:Person	schemaText	LITERAL	0-1
Website	schema:url	Base.gov.pt page	schema:Thing	schema:Url	IRI	0-1

Domain pproc:Contract subclassOf: pc:Contract						
Contract object name	schema:description	title of contract	schema:Thing	schema:Text	LITERAL	1
Procedure Type	pc:contractProcedure	how it was dealt	pproc:Contract	pc:procedureTypes	skos:Concept	0-n
type of contract	pc:kind	what was acquired	pproc:Contract	pproc:public-contracts-kinds	skos:Concept	0-n
Contract object	pc:mainObject	connection to further info	pproc:Contract	skos:Concept	skos:Concept	1
Buyer/Contractor	pc:contractingAuthority	buying company	pproc:Contract	gr:BusinessEntity	schema:Organization	1
Seller/Service Provider	pc:supplier	supplying entity	(not found)	(not found)	schema:Organization	1
Celebration Date	pc:awardDate	date when contract was closed	pproc:Contract	xsd:date	Schema:Date	0-1
End date	pc:actualEndDate	when contract was completed or extinct	pproc:Contract	xsd:date	Schema:Date	0-1
Working timeframe (days)	pproc:duration	days to fulfil contract	x	xsd:duration	LITERAL	0-1
Publication date	schema:datePublished	publication of contract notice (not contest)	schema:CreativeWork	schema:Date, schema:DateTime	Schema:Date	0-1
Place	pc:location	where was fulfilled	pproc:Contract	schema:Place	skos:Concept	0-n
Grounding	pproc:legalDocumentReference	Articles sustaining contract	pproc:Contract	x	LITERAL	0-1
Framework Agreement	pc:frameworkAgreement	is a framework agreement	pproc:Contract	pc:FrameworkAgreement	LITERAL	0-1
Contract Price	pc:agreedPrice	contracted price		gr:PriceSpecification	m4s:Euros	0-1
Actual price	pc:actualPrice	paid price	x	gr:PriceSpecification	m4s:Euros	0-1
Modifications to contract	pproc:contractModification	all modification reasons	pproc:Contract	pproc:ContractModification	bNode	0-1
Extinction of contract	pproc:contractOrProcedureExtinction	extinction information	pproc:Contract	pproc:ContractOrProcedureExtinction	bNode	0-1

domain pproc:ContractModification						
Why price Changed	pproc:modificationReason	Reasons for price shift	pproc:ContractModification	xsd:string	LITERAL	0-1
Why timeframe changed	pproc:durationChange	Reasons for time shift	pproc:ContractModification	xsd:string	LITERAL	0-1

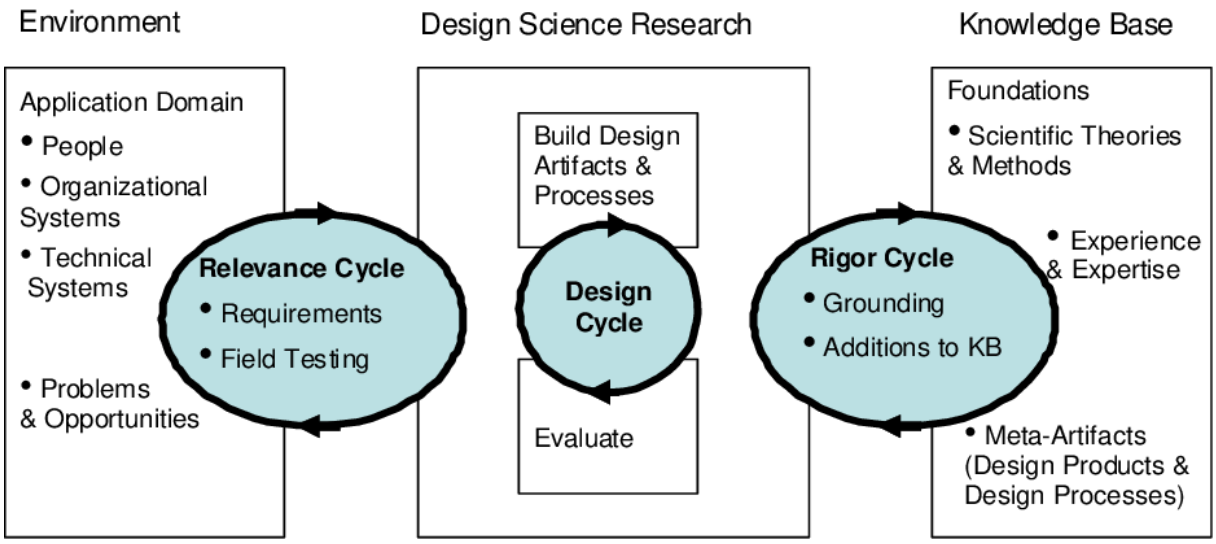
```
SELECT ?person ?organization ?nif ?role ?heads
WHERE {
   ?s <https://schema.org/memberOf> ?o .
    ?o schema:taxID  ?nif .
    ?o schema:location  ?heads .
    ?o schema:legalName  ?organization .
    ?s <https://schema.org/name> ?person .
    ?b rdf:subject ?s .
    ?b rdf:object ?o .
    ?b <https://schema.org/roleName> ?role .
} LIMIT 100

SELECT DISTINCT ?person ?organization ?nif ?role ?shares ?heads ?contract ?price
WHERE {
   ?s <https://schema.org/memberOf> ?o .
    ?o schema:legalName  ?organization .
    ?o schema:taxID  ?nif .
    ?o schema:location ?heads.
    ?s <https://schema.org/name> ?person .
    ?b rdf:subject ?s .
    ?b rdf:object ?o .
    ?b <https://schema.org/roleName> ?role .
    ?b <https://schema.org/description> "Acionista" .
    ?b m4s:companyShares ?shares .
    ?c pc:supplier ?o .
    ?c pc:actualPrice ?price .
    ?c schema:description ?contract .
} LIMIT 100


SELECT DISTINCT?s ?CPVlabel ?broadLabel ?sisterLabel
WHERE {

?s rdf:type pproc:Contract.
?s pc:mainObject ?CPV.
?CPV rdfs:label ?CPVlabel.
?CPV skos:broader ?broad.
?broad rdfs:label ?broadLabel.
?broad skos:narrower ?sister.
?sister rdfs:label ?sisterLabel
FILTER ( lang(?broadLabel) = "pt" ).
FILTER ( lang(?CPVlabel) = "pt" ).
FILTER ( lang(?sisterLabel) = "pt" ).

}LIMIT 100
```

```
:politician IRI  EXTRA a {
 a [m4s:Parliamentarian] ;
 schema:description LITERAL ;
 schema:url IRI ;
 schema:sameAs IRI *;
 schema:alternateName LITERAL ;
 schema:name LITERAL ;
 schema:gender [<https://schema.org/>~] ;
 schema:birthDate schema:Date ;
 schema:spouse LITERAL ? ;
 m4s:propertyRegimen [<https://example.org/map4scrutiny/value#>~]?;
 schema:hasCredential LITERAL OR <http://data.europa.eu/snb/eqf/>~]*;
 schema:hasOccupation LITERAL OR [<http://data.europa.eu/esco/>~] * ;
 m4s:parliamentaryComissions [<https://example.org/map4scrutiny/value#>~]*;
 dbo:parliamentaryGroup LITERAL ? ;
 schema:memberOf IRI @:organization + ;
 m4s:parliamentaryTerm LITERAL + ;
 m4s:constituency LITERAL OR [<http://vocab.getty.edu/tgn/>~] ;
 schema:comment LITERAL ? ; }

:organization IRI  EXTRA a {
 a [schema:Organization] ;
 schema:legalName LITERAL ;
 schema:alternateName LITERAL * ;
 schema:hasOfferCatalog LITERAL * ;
 schema:location [<http://vocab.getty.edu/tgn/>~] OR LITERAL * ;
 schema:taxID LITERAL ? ;
 schema:url IRI ? ; }

:contract IRI  EXTRA a {
 a [pproc:Contract] ;
 schema:description LITERAL ;
 pc:procedureType [<https://example.org/map4scrutiny/value#>~] ;
 pc:kind [<https://example.org/map4scrutiny/value#>~] OR LITERAL * ;
 pc:mainObject [<http://purl.org/cpv/2008/>~] + ;
 pc:contractingAuthority LITERAL OR IRI  @:organization ;
 pc:supplier LITERAL OR IRI @:organization ;
 pc:awardDate schema:Date ? ;
 pc:actualEndDate schema:Date ? ;
 pproc:duration LITERAL ;
 schema:datePublished schema:Date ;
 pc:location LITERAL OR [<http://vocab.getty.edu/tgn/>~] * ;
 pproc:legalDocumentReference LITERAL * ;
 pc:frameworkAgreement LITERAL ? ;
 pc:agreedPrice m4s:Euros ;
 pc:actualPrice m4s:Euros ? ;
 schema:url IRI ? ;
 pproc:contractModification BNODE @:modification ? ;
 pproc:contractOrProcedureExtinction BNODE @:modification ? ; }

:modification BNODE EXTRA a {
 a [pproc:ContractModification] ? ;
 a [pproc:ContractOrProcedureExtinction] ?  ;
 pproc:modificationReason LITERAL ? ;
 pproc:durationChange LITERAL ? ;
 pproc:extinctionCause LITERAL ? ; }

:metaRole BNODE EXTRA a {
  a [rdf:Statement] ;
  a [schema:Role] ;
  rdf:subject IRI ;
  rdf:predicate [schema:memberOf] ;
  rdf:object IRI ;
  schema:description LITERAL ?  ;
  schema:startDate schema:Date ?  ;
  schema:endDate schema:Date ?  ;
  schema:roleName LITERAL OR [<http://data.europa.eu/esco/>~] +  ;
  m4s:companyShares LITERAL ?  ;
  m4s:isAccumulation ["true"^^schema:Boolean "false"^^schema:Boolean] ; }
```

7 ATTACHMENTS

1	Supply oriented. The goal is to publish data in quantity.
2	Data needs to be requested or requires action to access. No API keys.
3	Values user engagement and feedback. Metadata emerges.
4	User incentive with hackathons and open communication channels.
5	Public interest, high-value data taxonomies are published
6	Linked Open Data and investment in digital skills development.
7	Data is high quality, timely and used for data-driven decisions.
8	Value creation and multiple sources working together with data communities and much more. An ecosystem is created, and data is no longer a goal but a means for a platform-based government.

Attachment 1 8 Steps to Mature Data according to the OECD [12]



Attachment 2 A Three Cycle View of Design Science Research [6]

8 REFERENCES

[1] K. Granickas, “Open Data as a Tool to Fight Corruption,” European Public Sector Information Platform *EPSI platform,* Topic Report No. 2014 / 04, 2014. Accessed: Feb. 24 2021. [Online]. Available: http://35.158.62.204/sites/default/files/2014_open_data_as_a_tool_to_fight_corruption.pdf

[2] J. Kim and M. Hausenblas, "Homepage: 5-star Open Data,". [Online]. Available: https://5stardata.info/en/ (Accessed: Jan. 27 2021).

[3] E. Folmer, S. Ronzhin, J. van Hillegersberg, W. Beek, and R. Lemmens, “Business Rationale for Linked Data at Governments: A Case Study at the Netherlands’ Kadaster Data Platform,” *IEEE Access*, vol. 8, pp. 70822–70835, 2020, doi: 10.1109/ACCESS.2020.2984691.

[4] E. Fagnoni, E. Norton, B. Acosta, M. Maleshkova, M. Domingue, J. Mikroyannidis, and A. Mulholland., “How to use Linked Data,” Linkeddata.center *wikitolearn.org*, Oct. 2020. Accessed: Feb. 24 2021. [Online]. Available: https://en.wikitolearn.org/Course:How_to_use_Linked_Data

[5] Transparency International, “Corruption Perception Index 2019,” Transparency International *Berlin, Germany*, 2020. Accessed: Feb. 24 2021. [Online]. Available: www.transparency.org/cpi

[6] A. Hevner, “A Three Cycle View of Design Science Research,” *Scandinavian Journal of Information Systems*, pp. 1–7, 2007. [Online]. Available: https://www.researchgate.net/publication/254804390

[7] Transparency - Corruptionary A-Z - Transparency.org,". [Online]. Available: https://www.transparency.org/en/corruptionary/transparency (Accessed: Nov. 1 2021).

[8] D. V. Malito, “Measuring Corruption Indicators and Indices,” Robert Schuman Centre for Advanced Studies *Italy,* EUI Working Paper RSCAS 2014/13, Feb. 2014. Accessed: Nov. 1 2021. [Online]. Available: http://hdl.handle.net/1814/29872

[9] Transparency International, "What is corruption?," *Transparency International.* [Online]. Available: https://www.transparency.org/en/what-is-corruption (Accessed: Jan. 28 2021).

[10] The World Bank, "Control of Corruption: Percentile Rank | Data Catalog," *World Bank,* 2020. [Online]. Available: https://datacatalog.worldbank.org/control-corruption-percentile-rank (Accessed: Jan. 28 2021).

[11] European Comission, “Report From The Commission To The Council And The European Parliament: Eu Anti-Corruption Report,” European Comission *Brussels,* COM(2014) 38 final, Feb. 2014. Accessed: Feb. 24 2021. [Online]. Available: https://ec.europa.eu/home-affairs/sites/

homeaffairs/files/e-library/documents/policies/organized-crime-and-human-trafficking/corruption/docs/acr_2014_en.pdf

[12] A. Vanroy, “European Semester Thematic Factsheet: Fight Against Corruption,” European Comission *ec.europa.eu*, Nov. 2017. Accessed: Feb. 24 2021. [Online]. Available: https://ec.europa.eu/info/sites/info/files/file_import/european-semester_thematic-factsheet_fight-against-corruption_en_0.pdf

[13] D. Kaufmann and A. Kraay, "Worldwide Governance Indicators: 1996–2019," *World Bank,* 2020. [Online]. Available: http://info.worldbank.org/governance/wgi/ (Accessed: Jan. 29 2021).

[14] Investopedia, "What Is Corporate Accountability?,". [Online]. Available: https://www.investopedia.com/terms/a/accountability.asp (Accessed: Nov. 1 2021).

[15] B. Welby, J. A. R. Perez, L. Chauvet, and G. Ugale, "The Path to Becoming a Data-Driven Public Sector,". Paris: OECD Publishing, 2019. Accessed: Feb. 24 2021. [Online]. Available: https://www.oecd-ilibrary.org/sites/059814a7-en/index.html?itemId=/content/publication/059814a7-en

[16] World Wide Web Foundation, “Open Data Barometer: Global Report,” World Wide Web Foundation *opendatabarometer.org*, 2017. Accessed: Feb. 24 2021. [Online]. Available: https://opendatabarometer.org/4thedition/report/

[17] J. Vrushi and R. Hodess, “Connecting The Dots:: Building the Case for Open Data to Fight Corruption,” Transparency International *www.transparency.org*, 2017. Accessed: Feb. 24 2021. [Online]. Available: https://images.transparencycdn.org/images/2017_OpenDataConnectingDots_EN.pdf

[18] D. Iglesias, “Open Data and The Fight Against Corruption in Brazil,” Transparency International *www.transparency.org* ISBN: 978-3-96076-039-9, 2017. Accessed: Feb. 24 2021. [Online]. Available: https://www.transparency.org/en/publications/open-data-and-the-fight-against-corruption-in-brazil

[19] A. G. Maail, “Linking Open Data and the Fight against Corruption in Indonesia,” *Perencanaan Pembangunan*, vol. 1, no. 3, pp. 256–264, Dec. 2017, doi: 10.36574/jpp.v1i3.23.

[20] A. Greco, “Open data and the fight against corruption in Latvia, Sweden and Finland,” Transparency International Latvia *delna.lv*, Nov. 2018. Accessed: Feb. 24 2021. [Online]. Available: https://delna.lv/en/2018/11/22/new-publication-open-data-and-the-fight-against-corruption-in-latvia-sweden-and-finland/

[21] J. A. R. Pérez and C. Emilsson, “OECD Open, Useful and Re-usable data: (OURdata) Index: 2019,” OECD, 2020. Accessed: Feb. 24 2021. [Online]. Available: http://www.oecd.org/gov/digital-government/ourdata-index-policy-paper-2020.pdf

[22] M. Moriconi and L. Bernardo, “Dados, Conhecimento, Ação:: Melhorar O Acesso À Informação Em Portugal,” Policy Paper SNI #1: Acesso à Informação *Transparência e Integridade*, 2012. Accessed: Feb. 24 2021. [Online]. Available: https://transparencia.pt/wp-content/uploads/2017/05/TIAC_SNI-AcessoInformacao2012.pdf

[23] World Wide Web Foundation, "Tech, Innovation and Open Government Convene At Open Up! 2012 Nov 13 In London," *Web Foundation,* 2012. [Online]. Available: https://webfoundation.org/2012/11/tech-innovation-and-open-government-convene-at-open-up-2012-nov-13-in-london/ (Accessed: Feb. 1 2021).

[24] N. Kroes, “Speech: The big data revolution,” Brussels, Mar. 26 2013. Accessed: Feb. 1 2021. [Online]. Available: https://ec.europa.eu/commission/presscorner/detail/en/speech_13_261

[25] M. Garcia, “Open Data and EU Funding,” European Public Sector Information Platform *EPSI platform,* Topic Report No. 2013 / 06, Jun. 2013. Accessed: Feb. 24 2021. [Online]. Available: https://www.europeandataportal.eu/sites/default/files/report/2013_open_data_and_eu_funding.pdf

[26] K. Barley and V. Jourová, “Relatório anual sobre o funcionamento do Registo de Transparência 2019,” Parlamento Europeu; Comissão Europeia, 2019. Accessed: Feb. 24 2021. [Online]. Available: https://www.europarl.europa.eu/at-your-service/files/transparency/pt-annual-report-on-the-operations-of-the-transparency-register-2019.pdf

[27] A. Lourenço, J. C. Ramalho, M. R. Gago, and P. Penteado, “Plataforma CLAV: contributo para a disponibilização de dados abertos da Administração Pública em Portugal,” *Cadernos BAD*, N.2, 19-44, 2019. [Online]. Available: https://bad.pt/publicacoes/index.php/cadernos/article/view/2047

[28] Agência para a Modernização Administrativa, I. P., "Portal de dados abertos da Administração Pública," *dados.gov.pt.* [Online]. Available: https://dados.gov.pt/pt/ (Accessed: Feb. 3 2021).

[29] Presidência da Modernização Administrativa, “Decreto-Lei n.º 83/2016,” in *Diário da República n.º 240/2016, Série I de 2016-12-16*, 2016, pp. 4728–4730. Accessed: Feb. 4 2021. [Online]. Available: https://data.dre.pt/eli/dec-lei/83/2016/12/16/p/

[30] Assembleia da República, “Lei n.º 3/2020,” in *Diário da República n.º 64/2020, Série I de 2020-03-31*, 2020, pp. 337–460. Accessed: Feb. 4 2021. [Online]. Available: https://data.dre.pt/eli/lei/3/2020/03/31/p/dre

[31] European Data Portal, "Introducing the new Open Data and PSI Directive," *europeandataportal.eu,* 2019. [Online]. Available: https://www.europeandataportal.eu/en/news/introducing-new-open-data-and-psi-directive (Accessed: Feb. 2 2021).

[32] M. Canares, K. Yusof, and S. Meng, "Collaborating For Open Data: Building An Open Database On Politically Exposed Persons In Malaysia:A Case Study," World Wide Web Foundation; Sinar Project, 2017. Accessed: Feb. 24 2021. [Online]. Available: http://webfoundation.org/docs/2017/08/RP-Collaboration-For-Open-Data-082017.pdf

[33] J. Florez and J. Tonn, "Accountability and anti-corruption,". Cape Town and Ottawa: African Minds and International Development Research Centre, 2019. Accessed: Feb. 17 2021. [Online]. Available: 10.5281/zenodo.2677862

[34] E. Bohórquez and R. G. Aceves, "Open Up Guide: Using Open Data to Combat Corruption," Open data charter, May. 2017. Accessed: Feb. 17 2021. [Online]. Available: https://open-data-charter.gitbook.io/open-up-guide-using-open-data-to-combat-corruption/

[35] World Economic Forum, "Partnering Against Corruption Initiative - Infrastructure & Urban Development: Building Foundations for Transparency," *Switzerland,* REF 180316, Mar. 2016. Accessed: Feb. 24 2021. [Online]. Available: https://www.weforum.org/reports/partnering-against-corruption-initiative-infrastructure-urban-development-building-foundations-for-transparency

[36] E. Parsons, "If you can't link to it... does it exist ?," *edparsons.com,* 2017. [Online]. Available: https://www.edparsons.com/2017/09/cant-link-exist/ (Accessed: Feb. 22 2021).

[37] C. Bizer, T. Heath, and T. Berners-Lee, "Linked Data - The Story So Far," *International Journal on Semantic Web and Information Systems*, vol. 5, no. 3, pp. 1–22, 2009, doi: 10.4018/jswis.2009081901.

[38] N. Ashish, P. Bhatt, and A. W. Toga, "Global Data Sharing in Alzheimer Disease Research," *Alzheimer disease and associated disorders*, vol. 30, no. 2, pp. 160–168, 2016, doi: 10.1097/WAD.0000000000000121.

[39] I. Jacobs, "Description of W3C Technology Stack," *w3.org,* 2010. [Online]. Available: https://www.w3.org/Consortium/techstack-desc.html (Accessed: Feb. 22 2021).

[40] D. Alami, I. Lera, C. Guerrero, and C. Juiz, "Experiences from the Design and Development of an Institutional Linked Open Data Portal," *TEM Journal*, vol. 6, no. 4, 2017, doi: 10.18421/TEM64-01.

[41] P.-Y. Vandenbussche and B. Vatant, "Homepage: Linked Open Vocabularies (LOV),". [Online]. Available: https://lov.linkeddata.es/dataset/lov/about (Accessed: Feb. 22 2021).

[42] OWL Working Group, "OWL - Semantic Web Standards," *World Wide Web Consortium,* 2012. [Online]. Available: https://www.w3.org/OWL/ (Accessed: Feb. 22 2021).

[43] A. Aggelen, L. Hollink, M. Kemman, M. Kleppe, and H. Beunders, "The debates of the European Parliament as Linked Open Data," *SW*, vol. 8, no. 2, pp. 271–281, 2016, doi: 10.3233/SW-160227.

[44] J. Hendler, J. Holm, C. Musialek, and G. Thomas, "US Government Linked Open Data: Semantic.data.gov," *IEEE Intell. Syst.*, vol. 27, no. 3, pp. 25–31, 2012, doi: 10.1109/MIS.2012.27.

[45] P. Espinoza-Arias, M. J. Fernández-Ruiz, V. Morlán-Plo, R. Notivol-Bezares, and O. Corcho, "The Zaragoza's Knowledge Graph: Open Data to Harness the City Knowledge," *Information*, vol. 11, no. 3, p. 129, 2020, doi: 10.3390/info11030129.

[46] Application Profiles Interest Group, "Dublin Core Tabular Application Profile (DCTAP)," *GitHub,* 2019. [Online]. Available: https://github.com/dcmi/DCTAP/ (Accessed: Dec. 11 2021).

[47] E. Prud'hommeaux, I. Boneva, J. L. Gayo, and G. Kellogg, "Shape Expressions Language 2.0," *W3C Community Final Specification Agreement (FSA),* 2019. [Online]. Available: https://shex.io/shex-primer/index.html (Accessed: Sep. 20 2021).

[48] J. M. Alvarez-Rodriguez, G. Alor-Hernández, J. E. L. Gayo, and C. Sanchez-Ramirez, "Towards A Pan-European E-Procurement Platform To Aggregate, Publish And Search Public Procurement Notices Powered By Linked Open Data: The Moldeas Approach," *Int. J. Soft. Eng. Knowl. Eng.*, vol. 22, no. 03, pp. 365–383, 2012, doi: 10.1142/S0218194012400086.

[49] A. Heise, F. Naumann, V. Ercegovac, and M. Hernandez, "GovWILD: Integrating Open Government Data for transparency," 2012, doi: 10.1145/2187980.2188039.

[50] C. Avila-Garzon, "Applications, Methodologies, and Technologies for Linked Open Data," *International Journal on Semantic Web and Information Systems*, vol. 16, no. 3, pp. 53–69, 2020, doi: 10.4018/IJSWIS.2020070104.

[51] J. P. McCrae, "Homepage: The Linked Open Data Cloud,". [Online]. Available: https://www.lod-cloud.net/ (Accessed: Feb. 22 2021).

[52] Publications Office Of The European Union, "Home - EU Vocabularies - Publications Office of the EU," *Publications Office Of The European Union.* [Online]. Available: https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies (Accessed: Nov. 17 2021).

[53] A. Drechsler and A. Hevner, Eds., "A four-cycle model of IS design science research,", Canada, 2016.

[54] Alan R. Hevner, Salvatore T. March, Jinsoo Park, Sudha Ram, “Design Science in Information Systems Research,” *MIS Quarterly*, vol. 28, no. 1, pp. 75–105, 2004. [Online]. Available: https://www.researchgate.net/publication/201168946

[55] Assembleia da República, “Lei 60/2019,” in *Série I, Diário da República Eletrónico*, 2019, 4 - 24. Accessed: May 13 2021. [Online]. Available: https://data.dre.pt/eli/lei/60/2019/08/13/p/dre

[56] Assembleia da República, "Sobre - Registo de Interesses," *https://www.parlamento.pt/.* [Online]. Available: https://www.parlamento.pt/RegistoInteresses/Paginas/default.aspx (Accessed: May 15 2021).

[57] S. Cook, C. Bock, P. Rivett, T. Rutt, E. Seidewitz, B. Selic, and D. Tolbert, "Unified Modeling Language, v2.5.1,", 2017. Accessed: Oct. 16 2021. [Online]. Available: https://www.omg.org/spec/UML/

[58] V. Krotov and L. Silva, Eds., "Legality and Ethics of Web Scraping,"Twenty-fourth Americas Conference on Information Systems, New Orleans, USA, 2018.

[59] D. Glez-Peña, A. Lourenço, H. López-Fernández, M. Reboiro-Jato, and F. Fdez-Riverola, “Web scraping technologies in an API world,” *Briefings in bioinformatics*, vol. 15, no. 5, pp. 788–797, 2014, doi: 10.1093/bib/bbt026.

[60] R. Verborgh and M. de Wilde, "Using OpenRefine,". Birmingham: Packt Publishing, 2013. Accessed: Oct. 18 2021. [Online]. Available: http://gbv.eblib.com/patron/FullRecord.aspx?p=1389316

[61] R. E. Mitchell, "Web scraping with Python,", 2nd ed. Sebastopol CA: O'Reilly Media, 2018.

[62] B. Schueler, S. Sizov, and S. Staab, “Management of Meta Knowledge for RDF Repositories,” in *International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA, USA, 092007, pp. 543–550.

[63] F. Erxleben, M. Günther, M. Krötzsch, J. Mendez, and D. Vrandečić, “Introducing Wikidata to the Linked Data Web,” in *Lecture Notes in Computer Science, The Semantic Web – ISWC 2014*, P. Mika et al., Eds., Cham: Springer International Publishing, 2014, pp. 50–65.

[64] Daniel Hernandez, Aidan Hogan, Markus Kroetzsch, “Reifying RDF: What Works Well With Wikidata?,” 2015. [Online]. Available: https://www.researchgate.net/publication/283865828

[65] A. Ismayilov, D. Kontokostas, S. Auer, J. Lehmann, and S. Hellmann, “Wikidata through the eyes of DBpedia,” *SW*, vol. 9, no. 4, pp. 493–503, 2018, doi: 10.3233/SW-170277.

[66] R. M. B. Kukutschka, “Global Corruption Barometer European Union 2021: Citizens' Views and Experiences of Corruption,” Transparency International, 2021. Accessed: Oct. 25 2021. [Online]. Available: https://www.transparency.org/en/publications/gcb-european-union-2021

[67] Karina, “Apagão do Portal BASE: o que está em causa?,” *Transparência Internacional Portugal*, 07 Oct., 2021. https://transparencia.pt/apagao-do-portal-base-o-que-esta-em-causa/ (Accessed: Oct. 18 2021).