DCPAPERS
Proc. Int'l Conf. on Dublin Core and Metadata Applications

Map4Scrutiny – A Linked Open Data Solution for Politicians Interest Registers

Inês Catarina Barreira Lopes
University of Minho,
Portugal
pg39356@alunos.uminho.pt

Ana Alice Batista
University of Minho,
Portugal
analice@dsi.uminho.pt

Óscar Atanázio Afonso
University of Porto,
Portugal
oafonso@fep.up.pt

Abstract

This article describes the process and results of providing data from the members of the Portuguese Parliament interest registers and the Public Procurement database as Linked Open Data. The organizations mentioned in both link the two resources. The two resources are linked by the organizations mentioned in both.

This article focuses on two areas: design and implementation. The first depicts the process of designing an application profile, structuring and describing the data, and creating shapes in ShEx for validation. Both existing and new properties, classes, and controlled vocabularies are used to describe the data in triples. During implementation, OpenRefine is used to clean and uniformize data, reconcile values into links, map triples from tabular to RDF, and export in RDF Turtle. The exported data is validated against the defined ShEx shapes, published in a Triplestore, and queried with SPARQL. The queries are used to showcase the difference between the sourced data and the resulting linked dataset.

This article describes every step of the implementation and provides a global vision of how to create Linked Open Data from structured and unstructured sources. These descriptions are particularly useful for researchers and professionals aiming to develop a similar project.
Keywords: design science research; linked open data; open government data; scrutiny; transparency

1. Introduction

Portugal is a country that scores high on the corruption perception index (Transparency International, 2020). Most Portuguese believe there is corruption in the country and are mostly concerned about Members of the Portuguese Parliament and Public Procurement (Bohórquez & Aceves, 2017). Background research supports the case for the use of Open Data (OD) as a corruption prevention tool (Granickas, 2014). However, in Portugal, OD scores low on impact and reuse (World Wide Web Foundation, 2017). Following the five-star deployment scheme, Linked Open Data (LOD) is the five-star form of OD (Kim & Hausenblas, 2018).

Therefore, the main motivators for this work are the public's interest in the subject and the potential for Open Data as a transparency prevention tool. Further research revealed a gap in the literature. There are OD projects and datasets considered to be corruption disablers, but few provide machine-readable and linked data, and even less provide LOD.

Seeing this gap in the Portuguese context led to the following research problem: 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 Open Data ready to be linked and reused by third parties.

The proposed solution for this problem consists of designing a LOD profile for the data, linking the two datasets to each other and existing LOD vocabularies, validating them, and uploading them at a Triplestore with an online SPARQL endpoint. SPARQL queries will then be used to illustrate the increased querying capability of LOD in comparison to the original formats.

https://doi.org/10.23106/dcmi.953176634

To answer a question as simple as: "How much did companies owned by parliamentarians made from public procurement in 2019?" Before this work, 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 analyzed to calculate 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, are published with LOD practices, this whole process is easily replaced by a SPARQL query.

This article is structured as follows: Section 2. Knowledge Base includes the prior art used to support the whole implementation; Sections 3 to 5 describe the methods used and clarify how the goal was achieved; Section 6 is the conclusions.

2. Knowledge Base

"Information that is not linkable is not used; information that is not used is not valuable" (Parsons, 2017). The solution proposed in this paper reuses data from www.base.gov.pt, a Portuguese public procurement portal considered a good example (European Comission, 2014), and the "Interest Registers" from politicians available in www.parlamento.pt. From the first, data can be retrieved in CSV, and from the second, in HTML.

Since there is no set methodology for implementing a LOD approach, the selected tools and the implementation steps are based on prior art. Particularly the work of Avila-Garzon (Avila-Garzon, 2020) that identified technologies used in several LOD management processes and the Linked Data Checklist from "How To Use Linked Data" (E. Fagnoni et al., 2020).

Also, because both selected sources are governmental, eGovernment LOD projects provided great insights: GovWILD - Integrating OGD for transparency (Heise et al., 2012); Data.gov - America's largest government LOD database (Kim & Hausenblas, 2018); LinkedEP - a LOD implementation describing European Parliament data in RDF (Aggelen et al., 2016); CLAV - a contribution to the availability of public administration OD in Portugal (Lourenço et al., 2019).

The implementation follows the four base principles for publishing LOD on the web defined by Tim Berners-Lee back in 2006: Use URIs as names for things; Use HTTP URIs so that people can look them up; Provide useful information on URIs using the standards; Include links to other URIs (Bizer et al., 2009).

To Encourage best practices, "How to use Linked Data" expanded these principles into a checklist: All relevant entities/concepts are extracted from the raw data; URIs are dereferenceable; Widely accepted vocabularies are used, and only non-existing terms are created; Dataset links to other RDF datasets; Created terms link to other vocabularies; Dataset includes metadata and information about licensing; There are alternative access methods: SPARQL endpoint, and Data dump; Dataset is registered in LD catalogs (E. Fagnoni et al., 2020).

An RDF Dataset is composed of triples. Each triple is structured resembling natural language. Subject – Predicate – Object. The subject is the asset being described (ex.: Subject A). The predicate states the relation between Subject and Object (ex.: rdf:type). The object answers to the predicate with the information on the subject (ex: Person). "Subject A has type Person" informs that the object described is a person. More predicates and objects are then added to the same subject to further describe it. For this to be linked data, the subject and predicate must be links. The object should, when possible and adequate, be a link, which means it can be the subject of another description.

In the examples above, the object Person is itself a subject with the rdf:type rdfs:Class. Classes in RDF are used precisely to describe the type of asset the resource is. Predicates are often called

https://doi.org/10.23106/dcmi.953176634

properties and, by convention, written in lower case (rdf:type) and classes written with a capital letter (rdfs:Class).

Selecting the appropriate vocabularies for properties, metadata schemes, and values' controlled vocabularies is one of the key points in designing a LOD profile. The EU is, for instance, 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 (Publications Office Of The European Union, 2021). A similar example is the UNESCO Thesaurus (E. Fagnoni et al., 2020).

There are standard, go-to schemes of properties and classes such as FOAF for describing the connection between people and Dublin Core Metadata Terms for describing resources. To find properties with natural language queries, there is Linked Open Vocabularies, a relevant search engine that should be used carefully because it does not include everything (E. Fagnoni et al., 2020) and (Avila-Garzon, 2020). Linked Data vocabularies catalogs are also useful for this purpose.

The properties in vocabularies often define a Domain and a Range. The first refers to the class of the subject the property is an instance of, and the second to the allowed values for that same property. These are instructions on how to link and describe instances of different classes. Failure to obey domains and ranges incurs a violation of the original vocabulary, in the sense that it goes against the intended and described the use of the original properties.

The selection of all the needed properties, classes and value vocabularies, and the relations between them are the map for the data transformation. To map this information in a structured way, the use of an application profile is advised. In this paper, two approaches are mentioned, the Constraint Matrix and DCMI's DCTAP application profile.

The Constraint Matrix is 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) in a table format. The first column has the local name of the property, and then the following columns have the link to the appropriate existing property, the description, the original domain, the original allowed values, and the cardinality (Malta & Baptista, 2013). The matrix is based on the Dublin Core Application Profile (DCTAP).

The DCTAP proposes a tabular format which should be filled bearing in mind the project's guidelines and the RDF and RDFS rules. This approach standardizes the profile development process in a simple and human-readable tabular format that can then be used to generate machine-readable validation schemas (Application Profiles Interest Group, 2019). Being an ongoing project naturally means it is not finished. However, this format is preferred to the Constraint Matrix because it is a format with guidelines for implementation available to the public, meaning that it can be interpreted by anyone. The matrix is kept anyway because it has extra information to help in an initial phase. This refers particularly to the original allowed values and domain that help avoid vocabulary violations.

OpenRefine with the RDF extension was used as the tool to apply the DCTAP to the data. It has powerful data cleaning functions and is particularly advised when handling tabular data (Bizer et al., 2009). Mapping to RDF with OpenRefine is applying the properties from the DCTAP to the columns in the table. OpenRefine also has a powerful reconciliation tool to turn values into links. This tool reads the uploaded value vocabularies, and then, for each column, the user selects the vocabulary in which OpenRefine should look for the values. OpenRefine compares the values of the column with the selected vocabulary 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. Finally, the tool offers the option to either directly publish on Wikidata or export a file in various formats (ex.: RDF-Turtle).

Before the RDF data was made available to the public, there was a shape 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 was built considering this future step using Shape Expressions (ShEX)

https://doi.org/10.23106/dcmi.953176634

as the validation language. Currently, an automated way to transform the DCTAP into ShEx is not available, but the ShEx schema equivalent to the DCTAP can be written manually with ShExC: A compact and very human-readable form of the language for describing Shapes Expressions with a syntax similar to RDF-Turtle.

The ShExC shapes have two main uses: To Aid future users in understanding how triples are shaped and for validation. Some samples of data profiling that are described with more detail in ShEx 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. Also, Open and Closed shapes can only be specified in ShEx. The difference is 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.

Validation is done by parsing the RDF data against the ShExC shapes. For this process, three files are needed, an RDF 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 Subjects A and B should conform to Shape 1, and Subjects C and D should pass as Shape 2 (Prud'hommeaux et al., 2019).

To run validation, one can use online or local instances. ShEx-java[1], available on GitHub, is one of the validators recommended in the ShEx documentation. It provides complete documentation and simple guidelines to the user.

For uploading the data and querying with SPARQL, the prior art shows OpenLink Virtuoso as one of the most used and stable triple stores, with SPARQL being the standard querying language (E. Fagnoni et al., 2020).

3. Methodology

The datasets on the interest registers and public contracts are linked according to the following rationale: The parliamentarians are the first piece to be described. Each is/was a member of none, one, 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 www.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 with a related subject.

The method applied to materialize the rationale above is described below with the aid of Activity Diagrams following the norms presented in Section 15 "Activities" of the Unified Modelling Language (UML) Specifications (Cook et al., 2017). Starting from the sourced data in a tabular format, Section 4 explains how the Linked Data Profile was designed, and Section 5 how the data was transformed to fit the designed profile, validated, uploaded to a Triplestore, and queried.

4. Design Linked Open Data Application Profile

This section describes the planning and designing work done before transforming the tabular data into LOD. This process includes the use of the Constrain Matrix, DCTAP, and ShEx Shapes.

The actions in FIG. 1 are based on the data retrieved from www.base.gov.pt and www.parlamento.pt. In tabular data, each column represents an attribute. Node 1 in FIG. 1 consists of selecting the attributes that will be kept and then listing them with "local" names to fill the first column "Local Properties".

https://doi.org/10.23106/dcmi.953176634

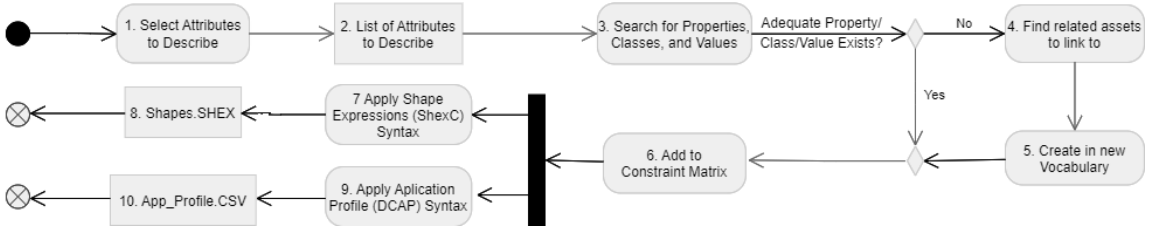


FIG. 1 Design a Linked Data Profile

a. Selecting LOD Vocabularies

Nodes 3 to 6 are crucial. These steps include the search for LOD vocabularies appropriate to link the data to other resources on the web of data. For every local attribute, an existing property with a URI should be found. The same is valid for controlled vocabularies and Classes.

Selecting the properties to re-use constituted one of the biggest challenges of implementation. There are many vocabularies 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.

The first approach to finding properties and classes focused on finding HTTP URIs in widely accepted, popular vocabularies (E. Fagnoni et al., 2020) and (Lourenço et al., 2019). The Italian Open Parliament Initiative[1] for describing parliamentarians, the Public Procurement Ontology[2] for public contracts, and the Organization Ontology[3] for organizations. Following the rules of these three ontologies for describing most of the data, plus further vocabularies when needed generate more triples than necessary to describe the data. This approach led to a larger dataset, and it would impact query ease of use and response time. Therefore, a second approach was used.

The second approach had three main goals: give priority to well-established property schemas, having a property with a parliamentarian as subject and an entity as an object or vice-versa, and use fewer vocabulary namespaces. If possible, the aim is also to have as few blank nodes as possible because "it is impossible to set external RDF links to a blank node, and merging data from different sources becomes much more difficult when blank nodes are used. Therefore, all resources of any importance should be named using URI references." (Bizer C. et al., 2007).

The main vocabulary used 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 by Google, Microsoft, and Yahoo! As a vocabulary to be read by search engines that would use it to provide richer results (E. Fagnoni et al., 2020).

This second attempt uses only 50 properties: 6 new resources, 9 namespaces, 6 classes, and 2 blank nodes. No data was discarded, the reduction of properties is due to the reduction in the number of classes needed because now more properties are directly connected to the main subjects.

Schema.org is used on 23 schema.org properties and replaced every instance of the organization ontology that was used on the first approach. The use of the Public Procurement Ontology and its predecessor, the Public Contracts Ontology, was maintained to describe the public contracts.

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 directly connected to the main subject as possible and reasonable. The two places for blank nodes are a compromise. One refers to the reified statements, and the second to modifications on the contracts.

1 http://dati.camera.it/ocd/reference_document/
2 http://contsem.unizar.es/def/sector-publico/pproc.html
3 https://www.w3.org/TR/vocab-org/

https://doi.org/10.23106/dcmi.953176634

b. Reification

After selecting the properties and classes to reuse one challenge remained: The representation of the connection between person and entity. To better explain this let us consider the following example:

The parliamentarian "John Doe" had the Role of consultant in "Society A" from 2015 to 2019.

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

John Doe – Subject

Member of – Predicate / Property

Society A – Object

However, to qualify this relation, 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 and can be represented in several ways in RDF (Schueler et al., 2007). Metaknowledge usually refers to information about data provenance, reliability, and timestamp. However, understanding its usability is complex and the same techniques used to describe it are also used for describing different data (Schueler et al., 2007).

Considering previous scientific work this prototype is going to use standard RDF reification. Literature shows it works, it has solid guidelines and a standard way to be described[4] (Ismayilov et al., 2018). Also, 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 (Daniel Hernandez, Aidan Hogan, Markus Kroetzsch, 2015). 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.

To aid in visualizing how the subjects are connected, the nodes representing a sample politician, organization, contract, and role, as described on the transformed data, are displayed in Attachment 1.

Even when prioritizing the use of known vocabularies, a satisfiable property or class for every local attribute was not always found. Either because no adequate resource was found, only Wikidata resources were found, or the URIs and descriptions were in were in Italian. The problem with Wikidata resources is that the URIs are number codes and therefore are not human readable.

This led to the creation of 6 new properties, 1 class, and 1 datatype. The new resources follow best practices and are connected to other existing vocabularies. The new resources also link to similar resources that were discarded. FIG. 2. Shows the example of the new property "Company shares" which refers to holding actions from a society. This is a type of ownership and for that reason schema:owns is an appropriate super-property. The included range is text because shares are displayed in varied formats such as the count of shares, percentage, or value.

```
: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
    .
```

FIG. 2 Extract from the Map4Scrutiny Vocabulary

In the end, to bring re-used and new resources together a single vocabulary file is created. Every re-used property is annotated in the vocabulary 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.

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

https://doi.org/10.23106/dcmi.953176634

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 IRIs, while some are strings, or when a different 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 attempting to reuse the resources have the basis to interpret whether their attributes 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 all new properties have as rdf:type only rdf:Property instead of being defined as either object or datatype properties to avoid excessive constraints. The vocabulary also includes a new datatype named "Euros" to be used in the price spaces to identify the currency.

c. Controlled Vocabularies

Having the properties and classes covered. The next step to completing the Constraint Matrix is defining what is expected as an object or value. For this purpose, both new and re-used controlled vocabularies of values are used. The first step was a search for controlled vocabularies for every value that seemed appropriate.

This search had two phases, first looking for controlled vocabularies in the European Union Publications Office and the UNESCO Thesaurus. Then do a free keyword search with standard web search engines for appropriate vocabularies. When the possible values for a property suited the creation of a controlled vocabulary, but nothing was found, a new vocabulary was created.

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

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

FIG. 3 Extract from the Map4Scrutiny Vocabulary of Values

FIG. 3 shows part of the conceptual schema for vocabularies used in the contract's procedure types and kinds. For both these top concepts, an approach already existed, and is linked in rdf:seeAlso. Because, the terms are not the same, and for legal reasons, only the clearly 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, classes, 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 SKOS 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, the base was the code from the EU Controlled Vocabularies, which were also reused by other projects (Alvarez-Rodriguez et al., 2012).

d. Dublin Core Tabular Application Profile and ShExC Shapes

As is shown in FIG. 1, once the Constraint Matrix was filled, it was formatted as a DCTAP application profile which was then used to manually write the shapes in ShExC. These files comprise all the decisions made above about properties, classes, and controlled vocabularies. In the scope of this project, everything could be described with the DCAP.

To showcase the similarities, a comparison between DCTAP and ShExC is shown in FIG. 4. The upper half is the DCTAP for the reified blank node_:, and the bottom half is the

https://doi.org/10.23106/dcmi.953176634

corresponding shape in ShExC. The content of the left box is almost equal in both formats, using "a" or rdf:type to indicate the relation between the instance and the class is the same thing.
Both models were kept because the DCTAP is more human-readable, which can help users understand easily what the data should look like. Also, it is helpful for researchers and professionals working on similar projects.

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] ;
```

FIG. 4 Extract from the Map4Scrutiny Shapes in ShExC and DCTAP - Roles Reified Statement

5. Transformation, Validationn, and Publishing

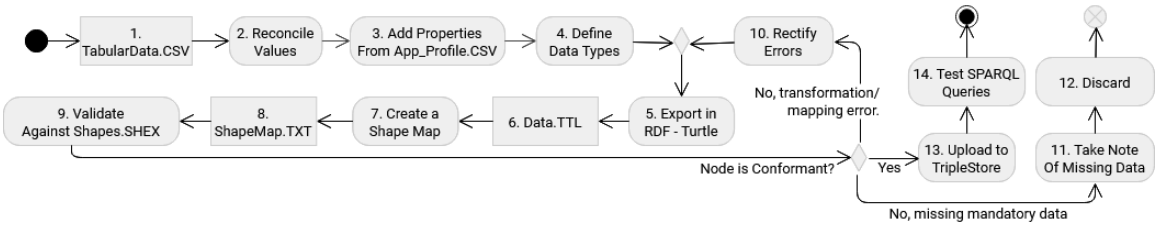


FIG. 5 Transform Validate and Publish

In FIG. 5, nodes 1 to 4 refer to activities that enabled the transformation of the tabular data into what is described in the DCTAP. They were all completed using OpenRefine.

Node 2 refers to using the reconciliation tool to transform string values into links. This was applied to a total of 10 columns with the names of geographical places, occupations, types of contracts, gender, and public procurement codes. From the 10 reconciled columns, only 3 kept at least one value as a string. In politicians' occupations, 194 were reconciled into an instance of the ESCO vocabulary (ex: Nurse > http://data.europa.eu/esco/isco/C2221) and 44 values remained as strings because either an equivalent was not found, or the string occupation did not have enough detail to be linked to the vocabulary. For instance: "Retired" or "Manager". The first is technically not an occupation, and the second is too broad to find a suitable link in the ESCO vocabulary. The type of contract values is also in a controlled vocabulary. However, 353 out of 133 067 types of contract values have a string that starts with "Other Type:" followed by a description of the specific case. These situations have no place in the controlled vocabulary. They are exceptions and are kept as strings.

With all the suitable values reconciled against controlled vocabularies, the next step was mapping the columns to the properties defined in the application profile. For each tabular file, the first column is the identifier that will be the subject, the following columns correspond to the object's properties, and the cells are the values. After identifying what column corresponds to what property, exporting the data to RDF-Turtle, node 5, was only one click away.

https://doi.org/10.23106/dcmi.953176634

a. Validation

The validation process is essential to make sure that the RDF-Turtle data is conformant to what is described in the application profile. In FIG. 5 it is represented by nodes 7 to 9. To run validation, a local implementation of ShEx-Java was used. The validator reads the shapes in ShExC, the RDF – Turtle data file, and the Shape Map.

The Reified Roles presented a challenge for validation because they are blank nodes and their shape identifier never appears as the object for a subject and property pair. When the chosen ShEx-Java validator parses the RDF-Turtle file, it gives blank nodes a new ID. So <_:node123> could after parsing be <_:nodeA>. The challenge of this characteristic is that when the validator reads a shape map stating that "<_:node123> @ <_:shape_reified>", _:node123 does not exist anymore and therefore cannot be validated. The solution to validate the reified statements against the defined shape was using an online validator[5] that accepts blank node IDs in its shape map and does not change the ID when parsing.

FIG. 4 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, and no triple was left unvalidated due to the use of two different validators.

In FIG. 5, Node 9 is followed by a decision node that depends on the output of the validation. Node 10 refers to "mapping and transformation errors". These are fixable mistakes made during transformation. One example is the Schema.org namespace that in the RDF-Turtle file started with "HTTP", and in the ShExC shapes with "HTTPS". This caused several nonconformant nodes that were easily fixed by correcting the namespace in the RDF-Turtle file. Another example was the whitespaces found in some of the Contracts' URIs. These were also fixed in OpenRefine and went back into the Nodes 5 to 9 sequence. When a correction could not be made the data was discarded, Nodes 11 and 12. The triples describing to the politicians and the entities all passed the validation process.

10 non-conformant reified roles were discarded because the original data had 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.

135 nonconformant contracts were discarded: 11 had no link to the supplier because the links were found using the unique fiscal identifier (NIF) and these suppliers had no NIF. 124 were discarded for having a common procurement vocabulary (CPV) code. The original shapes consider a link to the supplier to be mandatory because a link to the contractor and supplier is the minimum one could ask for when describing a contract as LOD. Also, because these are public contracts, having a CPV should also always be required.

b. Publication

Node 12 in FIG. 5 is uploading the data to a Triplestore. Following the recommendations in the literature, the validated RDF-Turtle data file was uploaded to a local instance of OpenLink Virtuoso to test SPARQL queries and verify if the subjects were properly linked to each other (Avila-Garzon, 2020), (E. Fagnoni et al., 2020), and (Bizer et al., 2009).

TABLE 1 Count of Triples in the Final Dataset

Types	Subjects	Triples	Vocabulary Name	Subjects	Triples
Parliamentarians	224	5 850	Map4Scrutiny New Properties	58	255
Organizations	25 519	101 866	Map4Scrutiny New Values	84	466

[5] https://shex.io/webapps/shex.js/doc/shex-simple.html

https://doi.org/10.23106/dcmi.953176634

Contracts	131 531	1 898 482	Getty Taxonomy Geographical Locations	30 341	357 697
Roles	1 650	14 942	ESCO – European Occupations Taxonomy	2 950	14 076
Contract Modification and Extinction	52 285	143 958	Common Procurement Vocabulary	10 420	298 583
Total	**211 209**	**2 165 098**	**Total**	**43 853**	**671 077**

TABLE 1 Count of Triples in the Final Dataset illustrates the size of the final dataset with a total of 2 165 098 triples describing entities, parliamentarians, and contracts. To make querying the data in a local implementation a friendlier experience, all the controlled vocabularies for values and vocabularies for properties and classes used to describe the data are also uploaded.

The SPARQL endpoint offered by OpenLink Virtuoso allowed running a set of queries both to verify if the data is linked properly and to 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 (TGN). Another example is the Common Procurement Vocabulary, where it is possible to navigate to narrower or broader connections of a given code. It is also now possible to query the data and get information from both datasets.

Going back to the question: "How much did companies owned by parliamentarians made from public procurement in 2019?" It can now be answered with a simple query.

TABLE 2 SPARQL Query Example

SPARQL Query

```
SELECT DISTINCT
( SUM ( ?price ) AS ?sum_price ) ( COUNT( ?c ) AS ?contract_count )
WHERE {
  ?s   rdf:type                m4s:Parliamentarian;
       schema:memberOf         ?o .

  ?b   rdf:subject             ?s ;
       rdf:object              ?o ;
       schema:description      "Acionista" .

  ?c   pc:supplier             ?o ;
       pc:actualPrice          ?value ;
       pc:awardDate            ?date;

  FILTER ( STRSTARTS ( STR ( ?date ), "2019" ) )
  BIND ( xsd:float ( STR ( ?value ) ) AS ?price ) }
```

Results

sum_price	contract_count
14950.04	1

Above in TABLE 2 SPARQL Query Example is a query that retrieves subjects that are Parliamentarians (?s) and members of an organization (?o). The parliamentarian and the organization are subject and object of a reified statement (?b) with the description "Acionista" which means that only companies where parliamentarians declare ownership of shares are considered. Then, the contracts (?c) where the organization was a supplier are queried for a date and price. The price of all contracts is then summed, and the number of contracts is counted. The transformed dataset only has one contract that is supplied by a company owned by a parliamentarian in 2019. It received 19 950 euros for the provided service. However, the same information can be queried for other timeframes. The same results can also be achieved by a similar query using the potential of the Datetime datatype.

6. Conclusions

The goal of this project was to create a sustained solution that follows LOD and Semantic Web guidelines and to document a full LOD transformation process. The final dataset is, by definition,

https://doi.org/10.23106/dcmi.953176634

five-stars OD because the data from both sources is connected to each other and to external data, such as the CPV, TGN, and ESCO vocabularies (Kim & Hausenblas, 2018). Further context is also provided with new controlled vocabularies created within the scope of the solution. Following the review of the objectives, the only task not yet completed due to external technical constraints is uploading the data to a server that maintains an online SPARQL endpoint for the public to explore. Currently, it is only available for bulk download[6].

The contribution to the knowledge base and domain affect 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[7]. These include a new Linked Data vocabulary of properties, classes, and a datatype properly connected to existing vocabularies, controlled vocabularies of values, 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, an LD description of the European Vocabulary for occupations, a DCTAP, the ShExC data shapes, and a description of the validation process together with the files used and needed to replicate process.

Further work includes translating the value vocabularies that are only available in Portuguese because they relate to legal matters and an official translation was not found and expanding the scope of the data by linking the organizations to other online profiles such as Wikidata and other related organizations.

References

Aggelen, A., Hollink, L., Kemman, M., Kleppe, M., & Beunders, H. (2016). The debates of the European Parliament as Linked Open Data. *Semantic Web*, *8*(2), 271–281. https://doi.org/10.3233/SW-160227

Alvarez-Rodriguez, J. M., Alor-Hernández, G., Gayo, J. E. L., & Sanchez-Ramirez, C. (2012). Towards A Pan-European E-Procurement Platform To Aggregate, Publish And Search Public Procurement Notices Powered By Linked Open Data: The Moldeas Approach. *International Journal of Software Engineering and Knowledge Engineering*, *22*(03), 365–383. https://doi.org/10.1142/S0218194012400086

Application Profiles Interest Group. (2019). *Dublin Core Application Profile (DCTAP)*. Dublin Core Metadata Initiative. https://github.com/dcmi/DCTAP/

Avila-Garzon, C. (2020). Applications, Methodologies, and Technologies for Linked Open Data. *International Journal on Semantic Web and Information Systems*, *16*(3), 53–69. https://doi.org/10.4018/IJSWIS.2020070104

Berners-Lee, T., & Bizer, C., Heath, T. (2009). Linked Data - The Story So Far. *International Journal on Semantic Web and Information Systems*, *5*(3), 1–22. https://doi.org/10.4018/jswis.2009081901

Bizer C., Cyganiak R., Heath T. (2007). How to publish linked data on the web. http://www4.wiwiss.fuberlin.de/bizer/pub/LinkedDataTutorial/

Bohórquez, E., & Aceves, R. G. (2017, May 17). *Open Up Guide: Using Open Data to Combat Corruption.* Open data charter. https://open-data-charter.gitbook.io/open-up-guide-using-open-data-to-combat-corruption/

Cook, S., Bock, C., Rivett, P., Rutt, T., Seidewitz, E., Selic, B., & Tolbert, D. (2017). *Unified Modeling Language, v2.5.1*. https://www.omg.org/spec/UML/

Daniel Hernandez, Aidan Hogan, Markus Kroetzsch (2015). Reifying RDF: What Works Well With Wikidata? https://www.researchgate.net/publication/283865828

E. Fagnoni, E. Norton, B. Acosta, M. Maleshkova, M. Domingue, J. Mikroyannidis, & A. Mulholland. (2020, October 23). *How to use Linked Data*. wikitolearn.org. Linkeddata.center. https://en.wikitolearn.org/Course:How_to_use_Linked_Data

European Comission. (2014, February 3). *Report From The Commission To The Council And The European Parliament: Eu Anti-Corruption Report* (COM(2014) 38 final). Brussels. European Comission. https://ec.europa.eu/home-affairs/sites/homeaffairs/files/e-library/documents/policies/organized-crime-and-human-trafficking/corruption/docs/acr_2014_en.pdf

[6] https://doi.org/10.34622/datarepositorium/K1DQIT
[7] https://doi.org/10.34622/datarepositorium/KWFSU2

https://doi.org/10.23106/dcmi.953176634

Granickas, K. (2014). *Open Data as a Tool to Fight Corruption* (Topic Report No. 2014 / 04). EPSI platform. European Public Sector Information Platform. http://35.158.62.204/sites/default/files/2014_open_data_as_a_tool_to_fight_corruption.pdf

Heise, A., Naumann, F., Ercegovac, V., & Hernandez, M. (2012). GovWILD: Integrating Open Government Data for transparency. Advance online publication. https://doi.org/10.1145/2187980.2188039

Hevner, A. (2007). A Three Cycle View of Design Science Research. *Scandinavian Journal of Information Systems*, 1–7. https://www.researchgate.net/publication/254804390

Ismayilov, A., Kontokostas, D., Auer, S., Lehmann, J., & Hellmann, S. (2018). Wikidata through the eyes of DBpedia. *Semantic Web*, *9*(4), 493–503. https://doi.org/10.3233/SW-170277

Kim, J., & Hausenblas, M. (2018, June 18). *Homepage: 5-star Open Data.* LOD Around The Clock. https://5stardata.info/en/

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

Malta, M. C., Baptista, A. A., & Center, A. (2013). A panoramic view on Metadata Application Profiles of the last decade-Matrix II Metadata Schemes.Parsons, E. (2017). If you can't link to it… does it exist ? edparsons.com. edparsons.com. https://www.edparsons.com/2017/09/cant-link-exist/

Prud'hommeaux, E., Boneva, I., Gayo, J. L., & Kellogg, G. (2019). *Shape Expressions Language 2.0: (ShEx) 2.1 Primer*. https://shex.io/shex-primer/index.html

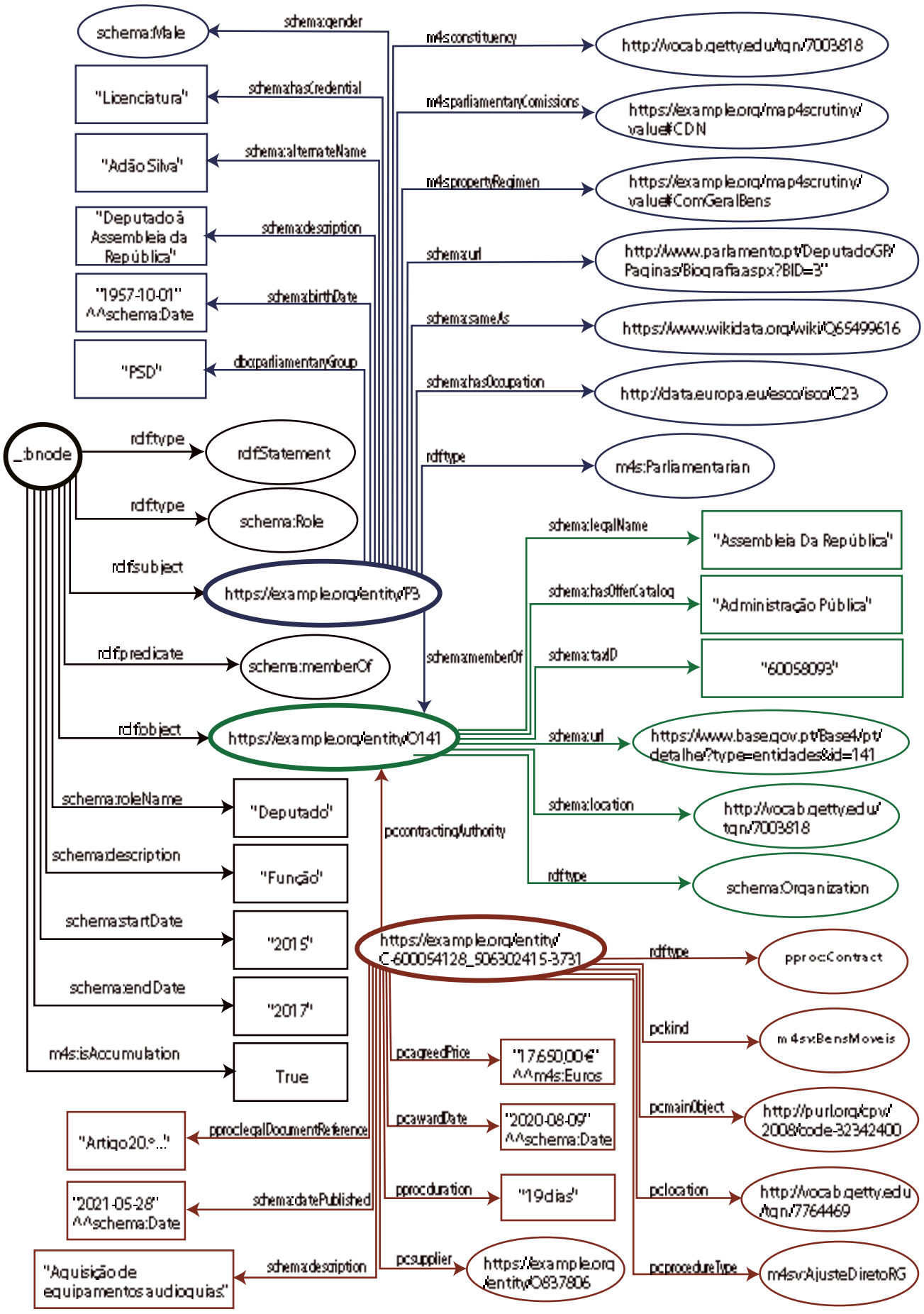
Publications Office Of The European Union. (2021, November 17). *Home - EU Vocabularies - Publications Office of the EU*. https://op.europa.eu/en/web/eu-vocabularies/controlled-vocabularies

Schueler, B., Sizov, S., & Staab, S. (092007). Management of Meta Knowledge for RDF Repositories. In *International Conference on Semantic Computing (ICSC 2007)* (pp. 543–550). IEEE. https://doi.org/10.1109/ICSC.2007.79

Transparency International. (2020). *Corruption Perception Index 2019*. Berlin, Germany. Transparency International. www.transparency.org/cpi

World Wide Web Foundation. (2017). *Open Data Barometer: Global Report*. opendatabarometer.org. World Wide Web Foundation. https://opendatabarometer.org/4thedition/report/

https://doi.org/10.23106/dcmi.953176634



Attachment 1 Sample Nodes From Transformed Data

https://doi.org/10.23106/dcmi.953176634