A Mapping Lifecycle for Public Procurement Data*

Eugeniu Costetchi[1,†], Alexandros Vassiliades[1,*,†] and Csongor I. Nyulas[1,†]

[1]Meaningfy SARL, 61 route de Fischbach, L-7447, Lintgen, Luxembourg

Abstract

To ensure the transparency of the public procurement procedures in the European Union (EU), procurement contracts above certain thresholds are governed by EU directives. Every year there are between 700,000 and 1.3 million public procurement related notices published on the Tenders Electronic Daily (TED) website. These notices are created based on the TED Standard Forms, and are published as XML data since 2014. In order to enhance the access and exploration of this data by a larger public, these notices are in the process of being transformed into RDF, conforming to the eProcurement Ontology (ePO). This poster illustrates the framework that we developed to cover the full lifecycle of creating modularised RML mappings to transform Public Procurement Data (PPD) from XML to RDF. The mapping creation lifecycle, has four phases that we repeat for each TED Standard Form: the creation of the conceptual mapping (CM), the creation of the technical mapping (TM), a validation phase, and the dissemination of the mapping. This poster also illustrates our innovative approach of creating reusable CM and TM modules, and automated validation queries, to ensure that our mappings generate a precise and complete RDF representation of the input.

Keywords

RDF Mapping Language (RML), Tenders Electronic Daily (TED), eProcurement Ontology (ePO), Conceptual Mapping, Technical Mapping

1. Introduction

Over 250,000 public authorities in the EU spend over €2 trillion (around 13.6% of GDP) yearly on the purchase of services, works and supplies. Procurement contracts above certain thresholds are governed by EU directives, to ensure the transparency of the procedures. Every year there are between 700,000 and 1.3 million public procurement related notices published on the Tenders Electronic Daily (TED) website[1]. These notices are published as XML data, since 2014, but in order to enhance the access and exploration of this data by the wider public they are in the process of being transformed into RDF data conforming to the structure of eProcurement Ontology (ePO)[2].

The RDF Mapping Language (RML)[3] offers a generic method, based on declarative rules, to

SEMANTiCS 2023: 19th International Conference on Semantic Systems, September 20–22, 2023, Leipzig, Germany

*Corresponding author.

eugen@meaningfy.ws (E. Costetchi); alexandros.vassiliadis@meaningfy.ws (A. Vassiliades); csongor.nyulas@meaningfy.ws (C. I. Nyulas)

https://meaningfy.ws/ (E. Costetchi)

0000-0002-9862-5070 (E. Costetchi); 0000-0003-4569-503X (A. Vassiliades)

[1]https://ted.europa.eu/TED/browse/browseByMap.do

[2]https://joinup.ec.europa.eu/collection/eprocurement/solution/eprocurement-ontology

[3]https://rml.io/specs/rml/

map data into an ontology while supporting various input data formats.

This poster presents the lifecycle of the mapping creation process, which is based on an innovative methodology to map European Union (EU) Public Procurement Data (PPD) published on the public TED website into ePO. At the core of the methodology is a process that we call the Conceptual Mapping (CM) of the data, which involves translating the numerous concepts that appear in each Standard Form into fragments of ePO. Then, based on this CM, we develop our RML mapping rules, a process we refer to as Technical Mapping (TM), which transforms data from XML files containing the encoded content of the completed Standard Forms into instances of ePO. In order to validate the quality of the data generated by the mapping process, we provide a validation method that automatically generates SHACL Data Shapes and SPARQL queries. Finally, if the quality of the produced RDF data is adequate, then the mapping is published, so that it can be used by an automatic processing pipeline. Currently the scalability of the method lies over the PPD presented in TED website, meaning that all data existing in that site in the form of Standard Forms or eForms[4] can be mapped. Theoretically, the method could over totally different contexts if the data are given in XML format, an appropriate CM is considered, and a TM developed.

The innovative methodology we provide for mapping PPD into ePO fragments is one of the main contribution we would like to illustrate through this poster. The work of converting PPD into ePO becomes even more difficult when you take into account the constant updating of PPD, such as the transition from Standard Forms to eForms, and the version updates of ePO that bring about changes in classes and relations. The TM then provides a general mapping methodology for applying RML mapping rules to map diverse PPD into the ePO ontology. In this mapping technique, we suggest that complexity be handled by managing the mapping rules as incomplete fragments, some of which are reusable and others of which are unique to a "mapping suite" (i.e., Form number). Another novelty of our approach is the validation method, which automatically generates SHACL Data Shapes and SPARQL queries to validate the accuracy of the provided data.

2. Mapping Methodology

In Figure 1, one can see the architecture of the framework presented in this paper. The pertinent PPD Standard Forms from the TED website are chosen for conceptual mapping in the *Conceptual Mapping* layer. In order to test and validate the mapping rules, a sample dataset is created. Additionally, a conceptual mapping is made by aligning business concepts, XML paths, and ontology fragments. The CM is then implemented using RML language in the layer called *Technical Mapping*, which is known as *Create Technical Mapping*. Additionally, the implemented TM rules change the sample dataset to enable quality checking. We provide SPARQL and SHACL validation procedures in the third layer (*Validation*), which assess the quality of the produced data. If violations and inconsistencies are discovered, the mechanism will indicate which areas of the CM appear to be problematic. Once the validation has been completed successfully, the mapping suite is made available (for a Notice Type) in the fourth layer *Dissemination* and is

[4]https://single-market-economy.ec.europa.eu/single-market/public-procurement/digital-procurement/eforms_en

stored in the mapping suite repository[5] to be used by the transformation pipeline as needed.

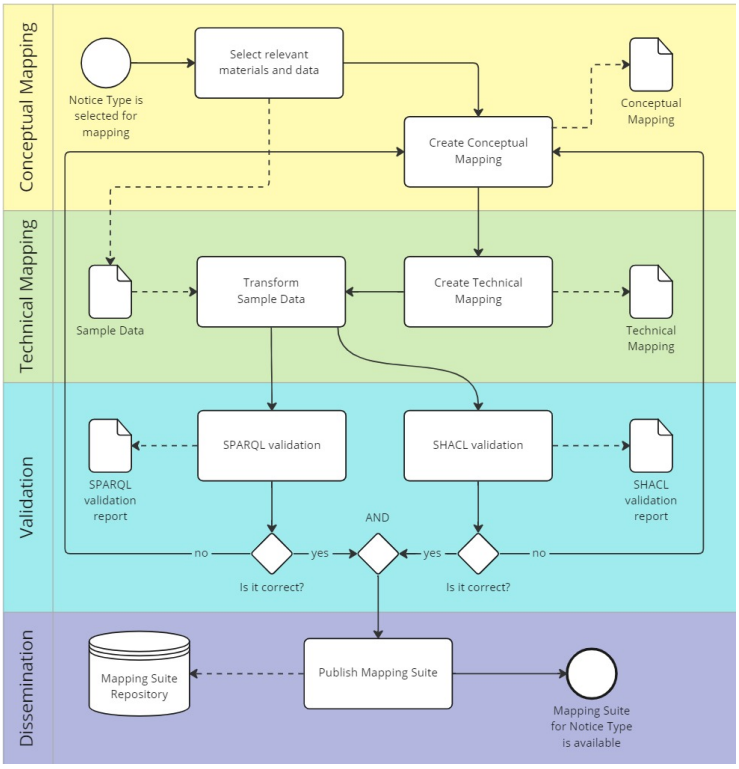

Figure 1: Mapping methodology workflow for the EU public procurement data

Nature of Source Data: The information that we are mapping into ePO refers to the PPD that is present in the TED's Standard Forms. These forms are available to assist citizens in publishing EU PPD in the EU Official Journal. For the purpose of disseminating this data, the European Commission has developed Standard Forms in accordance with each of the applicable EU legislative basis, namely: (i) TED schema forms set out in Regulation (EU) 2015/1986 and (ii) eForms set out in Regulation (EU) 2019/1780. More specifically, currently we mapped forms F03, F06, F13, F20, F21, F22, F23 and F25[6], and we will be progressing with the remaining ones.

Target Ontology - the eProcurement Ontology: The eProcurement Ontology (ePO) is a semantic data model that conceptualises and formally encodes the knowledge representation of the public procurement domain. Its primary purpose is to bridge the interoperability gap in the European public procurement data space, and can be used for data exchange, access and reuse.

Conceptual Mapping: The purpose of the CM is to translate the PPD Standard Form's sections, subsections, and fields into ePO ontology fragments—carefully selected collections of attributes and classes that accurately capture the instantiating context.

Technical Mapping: The term RML mapping mechanism refers to the declarative rules

[5]https://docs.ted.europa.eu/rdf-mapping/repository-structure.html

[6]see Standard Forms for Public Procurement (set out in Regulation (EU) 2015/1986) on SIMAP website: https://simap.ted.europa.eu/standard-forms-for-public-procurement

used to transfer the data from the XML files of the Standard Forms into RDF triples. However, these rules are only used to the extent that the toolchain allows and only to validate and test the data. The preliminary mapping we performed on our data made it easier for us to design the mapping rules because the CM clarified to which class and property we should map each element in the XML files.

We had to take into consideration some baselines for the RML mapping rules to be more configurable in order to handle the complexity of mapping the Standard Forms into ePO. We have applied the following solutions:

- *Sectioning within a form*, meaning that we have mappings for each form section in order to increase maintainability. When any changes apply to a section, rules for other sections will not be affected.
- *Segregation of rules* (generic and form specific), meaning that there are files that contain generic rules, reusable across mapping suites, and a file that is specific to a mapping suite.
- *Apply relative paths* in the mapping rules for handling versioning in the XML files.
- *Reuse of rules across Standard Forms and packages of Standard Forms*, meaning that there is a set of general source files where all the rules are kept as single source of truth. There is a selection and packaging process that picks the necessary modules to form a unified, self-sufficient package for each Standard Form.
- *Management of* rml:TripleMap *parts*, meaning that we had to separate the statements of *rml:subjectMap* and *rml:logicalSource* in form-specific modules, whereas the statements of *rml:predicateObject* are contained in modules reused across forms.

Dissemination: At the end of the mapping creation process, validated mapping suites are published in this GitHub repository https://github.com/OP-TED/ted-rdf-mapping/tree/main/mappings, from where they can be used to convert individual notices or integrated as part of a notice processing pipeline, such as the one provided here: https://github.com/OP-TED/ted-rdf-conversion-pipeline. The description of the GitHub repository where the mapping suites are published is documented at: https://docs.ted.europa.eu/rdf-mapping/index.html.

3. Evaluation

In this section, we analyze briefly the output from the SHACL and SPARQL validators, and discuss how to interpret the output to make the most of our mapping rules. Starting with the SHACL Data Shape validator, currently there are three types of violations, which can be categorized as (i) missing class relations (i.e., an instance is not correctly classified), (ii) cardinality constraints for more than one value, or (iii) cardinality constraints for less than one value.

In case of the SPARQL evaluation, we use a similar approach to list the types of inconsistencies that can occur, including unverifiable and invalid queries, warnings, and errors. Notice that the SPARQL queries presented in this section are not used for information retrieval for a user, they work as evaluation for the quality of the produced output, and are rather trivial SPARQL queries that check if data was mapped to a property.

The validation reports contain five result statuses: *Valid*, *Unverifiable*, *Warning*, *Invalid* and *Error*. Most of the results are *Valid* or *Unverifiable*, in case there is no input data in the sample to trigger a mapping rule. Some *Warning*s are signalled in cases when the field is found in the output, but not detected in the input. *Invalid* results are generated in cases when the data was found in the input, but is missing (or not detected by the current reporting tool) in the output. *Error*s occur when the query is wrong, or cannot be executed. No *Error*s are acceptable, and the few found in current reports are not real errors. A few *Invalid* results are found in the validation reports. Based on our analysis, they are not reflecting incorrect mapping rules or final data.

A total of 850 notices were processed by 1466 SPARQL queries that were automatically produced from the CM and distributed among 8 different kind of Standard Forms. More specifically, a set of 200, 195, 122, 146, 231, 231, 194 and 147 SPARQL queries were run, respectively, for each notice of the types F03, F06, F13, F20, F21, F22, and F25. Table 1 shows the number and percentage of queries for each type of inconsistency, over the total number of 217,179 query executions. The 82,477 query executions (or 37.98%), not shown in the table, were *Valid*.

Table 1
SPARQL Validator Result

Type of Inconsistency	Number of occurrence	Coverage
Error	151	0.07%
Invalid	3,988	1.84%
Unverifiable	104,860	48.28%
Warning	25,703	11.83%
Total	**217,179**	**62.02**%

4. Conclusion

This poster describes the lifecycle of creating RML mapping rules that can be used to convert PPD from the EU TED website into ePO, based on a novel mapping methodology. The first step is to map the various concepts of each Standard Form into fragments of ePO; this process is known as the CM of the data. Based on this CM, we developed our RML mapping rules, which convert the data from the XML files that the Standard Forms are represented in into instances of ePO classes; this process is known as the TM. To validate the accuracy of the generated data, we proposed a validation method that automatically generates SPARQL queries and SHACL Data Shapes. Finally, if the mapping package passes our validation process, we share it, to be used, independently or within conversion pipelines, to transform any XML notice data that was generated according to the Standard Form for which the mapping was developed into RDF.

We believe that applying the presented methodology carries many advantages. First off, having a CM makes it easier to create mapping rules, allowing for better "control" over where the data will be mapped, and allows for quality control of the output data. Next, our TM demonstrates how we might more effectively divide the data that needs to be mapped in order to modularise the mapping rules. Finally, by highlighting any areas where we need to modify a mapping rule, update the mapping we currently have in the CM, or fix a mapping rule, SPARQL and SHACL evaluators substantially ensure the quality of the provided data.