**Dr Krzysztof Węcel**
**[1]Uniwersytet Ekonomiczny w Poznaniu**
**[2]Instytut Informatyki Gospodarczej**
**krzysztof.wecel@ue.poznan.pl**

# Public Procurement in Linked Open Data Paradigm[1]

## 1. Introduction

The linked data paradigm[2] offers new opportunities with regard to publishing and re-using of data. Data is no longer rigidly defined and confined in tables but can easily be extended as new information in external sources is available. The new paradigm is particularly adopted by public institutions which have to fulfil requirements of transparency and open access to information[3]. This particular trend is called linked open data[4].

One of assets that can bring the greatest value both for enterprises and public bodies is data about public contracts. Contract notices, depending on their value, has to be published in national or European bulletins. In Poland, data is available in XML files, what makes a good case for transformation into a more flexible RDF format.

Enterprises are looking for appropriate contract opportunities using default search mechanism. Due to restrictions of interfaces, not all search

[1] This work was supported by a grant from the European Union's 7th Framework Programme provided for the project LOD2 Creating Knowledge out of Interlinked Data (GA no. 288176).
[2] Tim Berners-Lee, "Linked Data - Design Issues," *W3C* (ACM Press, 2006), 7 <doi:10.1145/1367497.1367760>.
[3] EU, "Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the Re-use of Public Sector Information," *Official Journal of the European Union*, 46 (2003), 90–96.
[4] Li Ding, Vassilios Peristeras and Michael Hausenblas, "Linked Open Government Data," *IEEE Intelligent Systems*, 27 (2012), 11–15 <doi:doi:10.1109/MIS.2012.56>.

criteria can be applied by external users interested in very specific contracts. Graphical interfaces are usually available in national languages, therefore access for foreign bidders is restricted. Furthermore, no wider analyses are available, like aggregations, trends, patterns discovery. Lack of mechanism for inferences rules prevents reasoning. Also, geographical data is not displayed or browsable via a map. Finally, external information cannot be easily integrated.

The above issues can be solved by sophisticated methods elaborated by semantic web and linked data communities[5].

## 2. Public Procurement Domain

Government procurement in European Union is regulated in several directives[6], then implemented by member states. Main principle is transparency, i.e. publication of calls for tenders is mandatory. Place of publication depends on the estimated value of a contract. Big contracts have to be advertised TED – Tenders Electronic Daily[7], the online version of the "Supplement to the Official Journal of the European Union", dedicated to European public procurement.

Theoretically, everybody has an access to data but not everybody is able to leverage it. The tendering process is not fully standardised among countries. Notices are available in various forms, e.g. HTML pages, DOC or PDF documents. Therefore, standardisation efforts emerged for representation of contract notices. One of such frameworks is Public Contracts Ontology (PCO[8]), presented in Figure 1.

---

[5] *Linking Enterprise Data* (Springer, 2010).
[6] e.g. Directive 2004/18/EC "on the coordination of procedures for the award of public works contracts, public supply contracts and public service contracts"
[7] http://ted.europa.eu
[8] https://code.google.com/p/public-contracts-ontology/

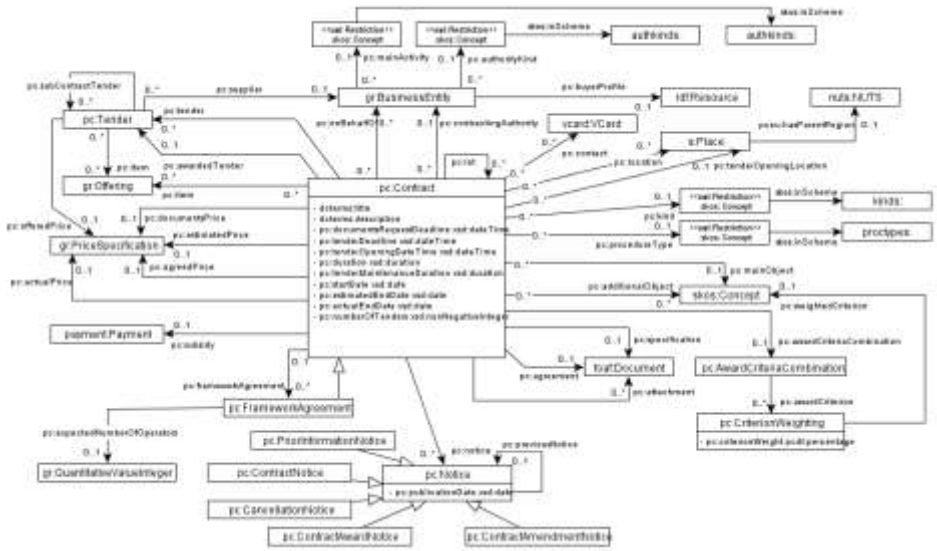

Figure 1 Public Contracts Ontology represented in UML Class Diagram

## 3. Polish Public Contracts

In Poland, public procurement data is published by The Public Procurement Office (Urząd Zamówień Publicznych[9]) in the Public Procurement Bulletin (Biuletyn Zamówień Publicznych[10]). Public procurement data is open to any person interested in using it and can be accessed in several ways: browsing the BZP portal, subscription mechanism with some restricted number of criteria or download of XML files[11]. The Office by offloading files to external server avoids overload of main server by people trying to scrap data from HTML pages.

Due to clarity of format we decided to work with XML files – one file per notice. They are published in daily updates packed into executable files to be run on Windows (.exe). Astonishingly, data required cleansing as some XML files contained illegal characters like `/u0x1`, `/u0x2`, `/u0x3`.

[9] http://uzp.gov.pl
[10] http://uzp.gov.pl/BZP/
[11] ftp://ftp.uzp.gov.pl/bzp/xml/

They had to be removed prior to using XML tools, otherwise the parser claimed that files were not valid XML. There were also some unescaped ampersand signs, causing confusion with XML entities.

There are nine document types published by Public Procurement Office, numbered from ZP-400 to ZP-408, with ZP-400 (contract notice) and ZP-403 (contract award notice) being the most popular. Document formats have common core structure and this has been leveraged by conversion scripts. There are just several extensions that cover specific requirements of notice. For example, in contract award notice value and price have to be given, while they were unknown in contract notice.

Listing 1 presents sample contract notice (ZP-400, shortened).

```
<?xml version="1.0" encoding="utf-8"?>
<ZP-400>
<biuletyn>1</biuletyn><pozycja>277559</pozycja>
<data_publikacji>2011-10-22</data_publikacji>
<nazwa>Komenda Powiatowa Państwowej Straży Pożarnej w Środzie
Wielkopolskiej</nazwa>
<ulica>ul. Libelta 2a</ulica>
<nr_domu>2a</nr_domu>
<nr_miesz></nr_miesz>
<miejscowosc>Środa Wielkopolska</miejscowosc>
<kod_poczt>63-000</kod_poczt>
<regon>63127679400000</regon>
<ogloszenie>ZP-400</ogloszenie>
<cpv1c>341140009</cpv1c>
<cpv2c>341130002</cpv2c>
...
</ZP-400>
```

Listing 1 Sample tender information in XML file

What is important to observe here is that the structure is basically flat. Even though some attributes can be grouped (e.g. an address) they

are put on the same level. It has implications for the parsing and conversion mechanisms. On one hand, no subset of XML data can be selected for further processing. On the other hand, the extraction expressions are shorter, as XML paths are shorter.

## 4. Conversion Process from XML to RDF

Conversion of XML files containing notices about public contracts has been carried out by means of tool Tripliser[12]. It is a powerful, flexible and relatively easy to use Java library and command-line tool for creating triple graphs from XML. Tripliser requires to think on semantic level, and therefore it is particularly suitable for data that is messy, bulky or volatile.

Several conceptual issues have been identified. First of all, *ZP-40x* document types have not been designed with easy processing in mind. One of the problems is a possibly unlimited list of tags, e.g. for `wykonawcy` (executors) and `cz` (parts) consecutive numbers are used: `wykonawca_0`, `wykonawca_1`, `cz_0`, `cz_1` and so on (see Listing 2).

```
<czesci>
  <cz_0>
    <wykonawcy>
      <wykonawca_0>
        <nazwa_wyk>Katarzyna Rżanek</nazwa_wyk>
        <adres>ul. Karwińska 22</adres>
        <miejsc>Warszawa</miejsc>
        <kod>02-639</kod>
        <wojewodztwo>mazowieckie</wojewodztwo>
      </wykonawca_0>
      <wykonawca_1>
        <nazwa_wyk>Wojciech Gola</nazwa_wyk>
        ...
```

[12] http://daverog.github.io/tripliser/

```
      </wykonawca_1>
```

Listing 2 Tripliser: mapping of executor to parts of a contract

Such XML required specific approach to extraction. It was not possible to use contract as a subject and property `pc:supplier` to assign executor. Instead, we had to start with an executor as a subject and then query for the contract. Consequently, a new inverse property `pc:supplierFor` had to be defined. Unfortunately, it is not possible to define property as inverse in Tripliser mapping file.

Another challenge that has made mapping files more complex was creation of identifiers for resources. Our first naive approach was to use blank nodes for representation of nested information like for example addresses. Listing 3 present this initial approach.

```
<http://data.i2g.pl/res/vcard#43120573100000>
  a     v:VCard , v:Work ;
  v:adr [ a       v:Address , v:Work ;
    v:country-name "Poland" ;
    v:extended-address "ul. Bema 1 1 m. " ;
    v:locality "Puławy" ;
    v:postal-code "24-100" ;
    v:region "lubelskie" ;
    v:street-address "ul. Bema 1"
  ].
```

Listing 3 Information about organisation in turtle format

After loading all files we quickly run into problems – there were the same addresses, faxes and phone numbers attached many times to a single organisation. So many times as there were number of published notices by them. The reason is that blank nodes are separate entities and they do not merge. We had to provide an identifier for each resource produced from XML file. This should be remembered as a guideline for each triplification process. Capabilities of XPath were not sufficient to generate

unique identifier. Therefore, we implemented own extension functions in Java to be used in Tripliser, addressing specific naming conventions and limitations. Identifier for address is generated from its parts: locality, postal code and street. The function `fn:encode-address` returns part of city name followed by hashed elements of address (e.g. `pula-80bd-c14d-4168ce0b` in Listing 4).

Similar approach was applied for executors – natural person without any official number can also be an awarder executor. Such a solution is perfect for merging of information from various notices or even external datasets.

```
<http://data.i2g.pl/zp/vcard/43120573100000>
  a      v:VCard ;
  v:adr  <http://data.i2g.pl/zp/address/pula-80bd-c14d-4168ce0b> ;
  v:email <mailto:jkowalczyk@poczta.pulawy.pl> ;
  v:fax  "081 8877023" ;
  v:fn   "Samodzielny Publiczny Zakład Opieki Zdrowotnej" ;
  v:tel  "081 8877023" ;
  v:url  "spzoz.n2.pl" .
<http://data.i2g.pl/zp/address/pula-80bd-c14d-4168ce0b>
  a       v:Address ;
  v:country-name "Poland" ;
  v:extended-address "ul. Bema 1 1 m. " ;
  v:locality "Puławy" ;
  v:postal-code "24-100" ;
  v:region "lubelskie" ;
  v:street-address "ul. Bema 1" .
```

Listing 4 Information about organisation in turtle format - refined approach

An important part of XML to RDF mapping was preparation of mapping files. The mapping language is quite straightforward provided that one is familiar with XPath. Query within a `<graph>` is used to select all XML elements from which further translation will take place. Element

`<resource>` is used to create RDF resource with respective properties. The resource needs an identifier, which is defined within `<about>`. Properties are contained in `<property>`, and can be nested when in need. Values can be assigned to properties statically using `value` attribute or dynamically by means of `query` attribute. A fragment for extraction of executor is presented in Listing 5.

```
<resource query="czesci/*/wykonawcy/*" comment="Wykonawcy">
 <about prepend="${org}" query="fn:encode-for-uri(myfn:encode-executor(
   nazwa_wyk,miejsc,adres))" required="true" />
 <properties>
  <property name="pc:contact" resource="true" prepend="${vcard}" query=
    "fn:encode-for-uri(myfn:encode-executor(nazwa_wyk,miejsc,adres))" />
  <property name="gr:legalName" query="nazwa_wyk" />
  <property name="rdf:type" resource="true" value="gr:BusinessEntity" />
  <property name="rdf:type" resource="true" value="zp:Wykonawca" />
  <property name="pc:supplierFor" resource="true" prepend="${tender}"
    required="true" query="concat(fn:year-from-date(//data_publikacji),
    '_', //pozycja, '_', parent::*/parent::*/nr_czesci_1)" />
 </properties>
</resource>
```

Listing 5 Tripliser: mapping of executor to parts of a contract

The fragment presents some sophisticated constructs that had to be used in order to extract data from XML: wildcard is used to match unknown tags; built-in functions `fn:encode-for-uri`, `fn:year-from-date` as well as own extension function `myfn:encode-executor` are used, finally advanced XPath expressions `parent::*/parent::*/` to navigate back in hierarchy.

It is important to emphasize that our resulting RDF model is consistent with Public Procurement Ontology (PCO). We also accepted and extended its naming convention – reflected in prefixes and identifiers of resources. Listing 6 presents a sample contract extracted using the above

procedure. According to PCO, parts of the contract are also modelled as contract but they use additional attribute `pc:isLotFor` to refer to main contract. As an output formats we prefer Turtle which is much more concise and better readable for humans than RDF/XML serialisation.

```
@prefix br:      <http://purl.org/business-register#> .
@prefix v:       <http://www.w3.org/2006/vcard/> .
@prefix pc:      <http://purl.org/procurement/public-contracts#> .
@prefix zp:      <http://i2g.pl/voc/zamowienia-publiczne#> .
@prefix gr:      <http://purl.org/goodrelations/v1#> .
<http://data.i2g.pl/zp/contract/2012_294490>
  a       pc:Contract ;
  zp:rodzaj_zam "Inny: Państwowa Jednostka Organizacyjna" ;
  dc:description """Zadanie nr 1  ...""" ;
  dc:name "Opracowanie dokumentacji techniczno-projektowej budowy dróg
  leśnych" ;
  pc:additionalObject
<http://purl.org/weso/pscs/cpv/2008/resource/713200007> ;
  pc:contact <http://data.i2g.pl/zp/vcard/65001711200000> ;
  pc:contractingAuthority
  <http://data.i2g.pl/zp/organization/65001711200000> ;
  pc:kind kinds:Services ;
  pc:mainObject <http://purl.org/weso/pscs/cpv/2008/resource/452331206>;
  pc:notice <http://data.i2g.pl/zp/notice/2012_294490> ;
  pc:procedureType proctypes:Open ;
  pc:publicationDate "2012-08-09" ;
  pc:referenceNumber "2012_294490" .
<http://data.i2g.pl/zp/contract/2012_294490_1>
  a       pc:Contract ;
  zp:data_zam "2012-08-08"^^xsd:date ;
  zp:liczba_odrzuconych_ofert "0"^^xsd:int ;
  zp:nr_czesci "1" ;
  dcterms:name "Opracowanie dokumentacji projektowej budowy drogi leśnej
  Cisowa - Łodzinka - Panieński Czub w Leśnictwie  Cisowa o dług. L - 5
  550,0m." ;
```

```
pc:awardedTender <http://data.i2g.pl/zp/tender/2012_294490_1> ;
pc:estimatedPrice <http://data.i2g.pl/zp/price/2012_294490_1-est> ;
pc:isLotFor <http://data.i2g.pl/zp/contract/2012_294490> ;
pc:item <http://data.i2g.pl/zp/offering/2012_294490_1> ;
pc:numberOfTenders "7"^^xsd:int .
```

Listing 6 Contract award notice with parts in Turtle format

There were altogether 404,713 notices published in 2012. We have converted 179,205 notices about awarded contracts, posted by 15,968 contracting authorities, resulting in over 14 million triples. Single file was loaded in 1.8 sec. on average, and total load time 1h 40min. Table 1 summarises the conversion process.

Table 1 Summary statistics for public procurement conversion

| | |
|---|---|
| Total number of notices in 2012 | 404,713 |
| Average number of notices per month | 33,726 |
| Number of ZP-400 documents | 174,225 |
| Number of ZP-403 documents | 179,205 |
| Number of XML files converted per second | 300-400 |
| Duration of one month conversion | ca. 80-110 sec. |
| The whole year conversion | ca. 20 min. |
| Time for Virtuoso upload | 1h 40 min. |

Source: own calculations

## 5. Enrichment of Public Contracts Data

One of the advantages of representing data in RDF mentioned in introduction was the possibility to link to external facts. For example, products can be represented by product ontologies[13], and executors can be linked to external registries[14]. Also, external documents can be included, e.g. objections, results of judicial decisions, possible NLP analysis of Terms of Reference (SIWZ).

Our requirement was to visualise public contract statistics on the map with drill-down functionality. In public procurement data, addresses

[13] e.g. GoodRelations, http://www.heppnetz.de/projects/goodrelations/
[14] e.g. OpenCorporates, http://opencorporates.com/

do not contain information about *powiat* (district). It can be inferred but requires additional search in other databases. Our maps of Poland have units identified by TERYT number. Therefore, it was necessary to map contact of given contracting authority or executor to TERYT, enriching our dataset with geographical dimension. For the purpose of mapping we utilised one of LOD2 stack tools – SILK[15]. The tool identified 22,076 links between addresses and TERYT localities. After geocoding, it was possible to prepare choropleth maps as presented in Figure 2.

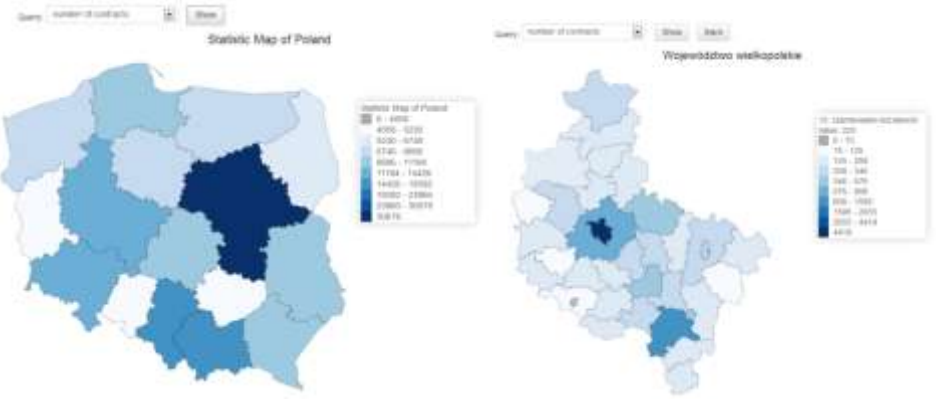


Figure 2 Number of public contracts presented on cuntry and districts maps

## 6. Conclusions

As a result of conversion process all data on public procurement along with supplementing and extensible data is readily available in a convenient format and can be uniformly queried with SPARQL. Analysts can prepare the reports that they need, as pictured on maps above, supporting better decision. It is also possible to leverage the standardised interfaces and applications developed for the domain, e.g. filing and matching application developed by colleagues from Czech Republic. The

[15] Robert Isele, Anja Jentzsch and Christian Bizer, "Silk Server - Adding Missing Links While Consuming Linked Data," in *Proc. of 1st International Workshop on Consuming Linked Data (COLD 2010)* (Shanghai, 2010).

approach can much easier not only assure transparency so demanded by democracy but also increase the efficiency of economy overall. Several groups of beneficiaries can profit: contracting authorities – better prices expectations, more precise offers; companies – comparison of prices, better targeting, information about competitors; supervising authorities – transparency, access to controlling tools; and finally citizens – more efficient use of public money.

## Bibliografia


Berners-Lee, Tim, "Linked Data - Design Issues," *W3C* (ACM Press, 2006), 7 <doi:10.1145/1367497.1367760>

Ding, Li, Vassilios Peristeras, and Michael Hausenblas, "Linked Open Government Data," *IEEE Intelligent Systems*, 27 (2012), 11–15 <doi:doi:10.1109/MIS.2012.56>

EU, "Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the Re-use of Public Sector Information," *Official Journal of the European Union*, 46 (2003), 90–96

Isele, Robert, Anja Jentzsch, and Christian Bizer, "Silk Server - Adding Missing Links While Consuming Linked Data," in *Proc. of 1st International Workshop on Consuming Linked Data (COLD 2010)* (Shanghai, 2010)

Wood, David, ed., *Linking Enterprise Data* (Springer, 2010)


## Summary


The paper presents a conversion process of public contracts notices from XML files to RDF triples. The linked data approach has several advantages over classical databases: navigation between data, and enrichment of possessed datasets by linking to external information. It allows new application areas for enterprise like semantic search for offers and precise matching of tenders to interested contractors. Specific vocabulary has also been developed to make integration possible.