Removing Barriers to Transparency: a Case Study on the Use of Semantic Technologies to Tackle Procurement Data Inconsistency

Giuseppe Futia*, Alessio Melandri**, Antonio Vetrò*, Federico Morando**, Juan Carlos De Martin*

* Nexa Center for Internet and Society, Politecnico di Torino (DAUIN)
** Synapta Srl

Outline

- Public procurement context
- Our contribution
- Results
- Discussion
- Future works

Context

What is Transparency?

Data related to functioning of government can be **accessed** and **interpreted**, without being **pre-processed** and **manipulated**

Open Government Data (OGD) is (or should be) a mechanism integrated in the heart of the government functions for creating transparency

Public Procurement (PP) is a specific area of the OGD that increases **openness of government information** and supports **business activities**

Is OGD Enough for Enabling Transparency through Procurement Data?

- An obstacle to transparency is related to the **fragmentation of OGD** in the domain of procurement data

- Fragmentation can generate **discrepancies and inconsistencies** among data that can not be a priori detected and fixed

The Italian Context

Our case study is represented by **Italian PP data** published on different websites of Italian administrations - in compliance with the **Italian anti-corruption Act** (law n. 190/2012)

```
▼<lotto>
  <cig>Z380F6F11</cig>
 ▼<strutturaProponente>
    <codiceFiscaleProp>00518460019</codiceFiscaleProp>
    <denominazione>Politecnico di Torino</denominazione>
  </strutturaProponente>
 ▼<oggetto>
    indagini idrogeologiche per la realizzazione di pozzi geometrici collegati a pompe di calore
  </oggetto>
  <sceltaContraente>08-AFFIDAMENTO IN ECONOMIA - COTTIMO FIDUCIARIO</sceltaContraente>
 ▼<partecipanti>
  ▼<partecipante>
    <codiceFiscale>FLRLCU75C23F335C</codiceFiscale>
    <ragioneSociale>STUDIO APOGEO (DOTT. LUCA FILIERI)</ragioneSociale>
   </partecipante>
  </partecipanti>
 ▼<aggiudicatari>
  ▼<aggiudicatario>
    <codiceFiscale>FLRLCU75C23F335C</codiceFiscale>
    <ragioneSociale>STUDIO APOGEO (DOTT. LUCA FILIERI)</ragioneSociale>
   </aggiudicatario>
  </aggiudicatari>
  <importoAggiudicazione>5000.00</importoAggiudicazione>
 ▼<tempiCompletamento>
    <dataInizio>2013-10-31</dataInizio>
  </tempiCompletamento>
  <importoSommeLiquidate>00.00</importoSommeLiquidate>
</lotto>
```

Table 1. Number of downloaded XML files of procurement data in different periods of time

	May 2015	Nov 2015	Feb 2016	Nov 2016
URL requested	-	-	207.674	271.664
downloaded files	205.415	184.738	201.451	252.246
valid XMLs	199.341	180.609	197.338	247.881

300.000 XML files from
more than **16.000** public bodies

What is Data Consistency?

Data consistency refers to **the absence of apparent contradictions** within data (ISO/IEC 25012)

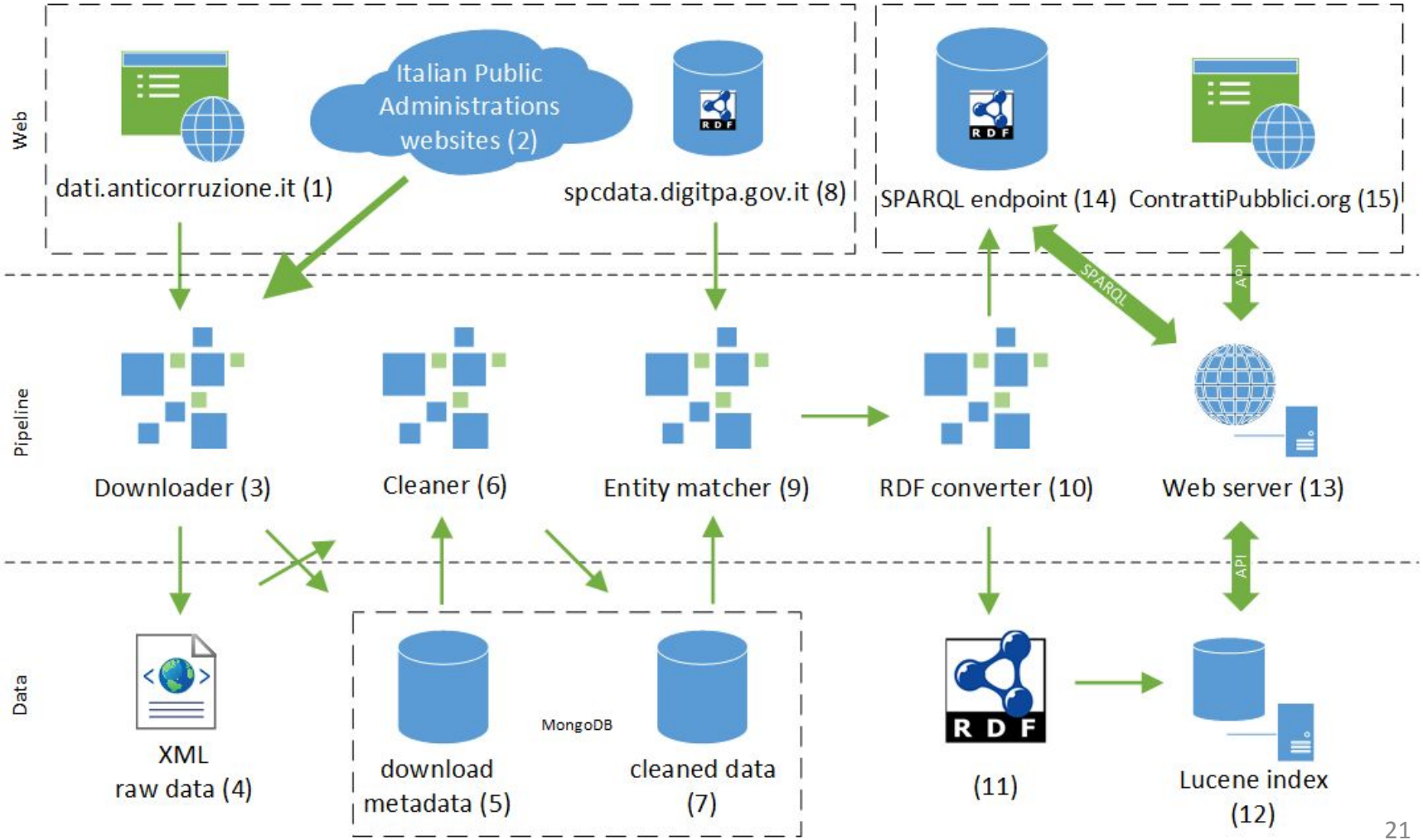
- Certain types of consistency problems directly emerge analyzing contracts **data collected in a single XML file**

  - contracts in which **the beneficiary is more than one**
  - contracts in which the amount of money is paid, but **no recipient is present in the data**
  - contracts in which the **sum reported as paid is greater than the sum initially awarded** to the beneficiary

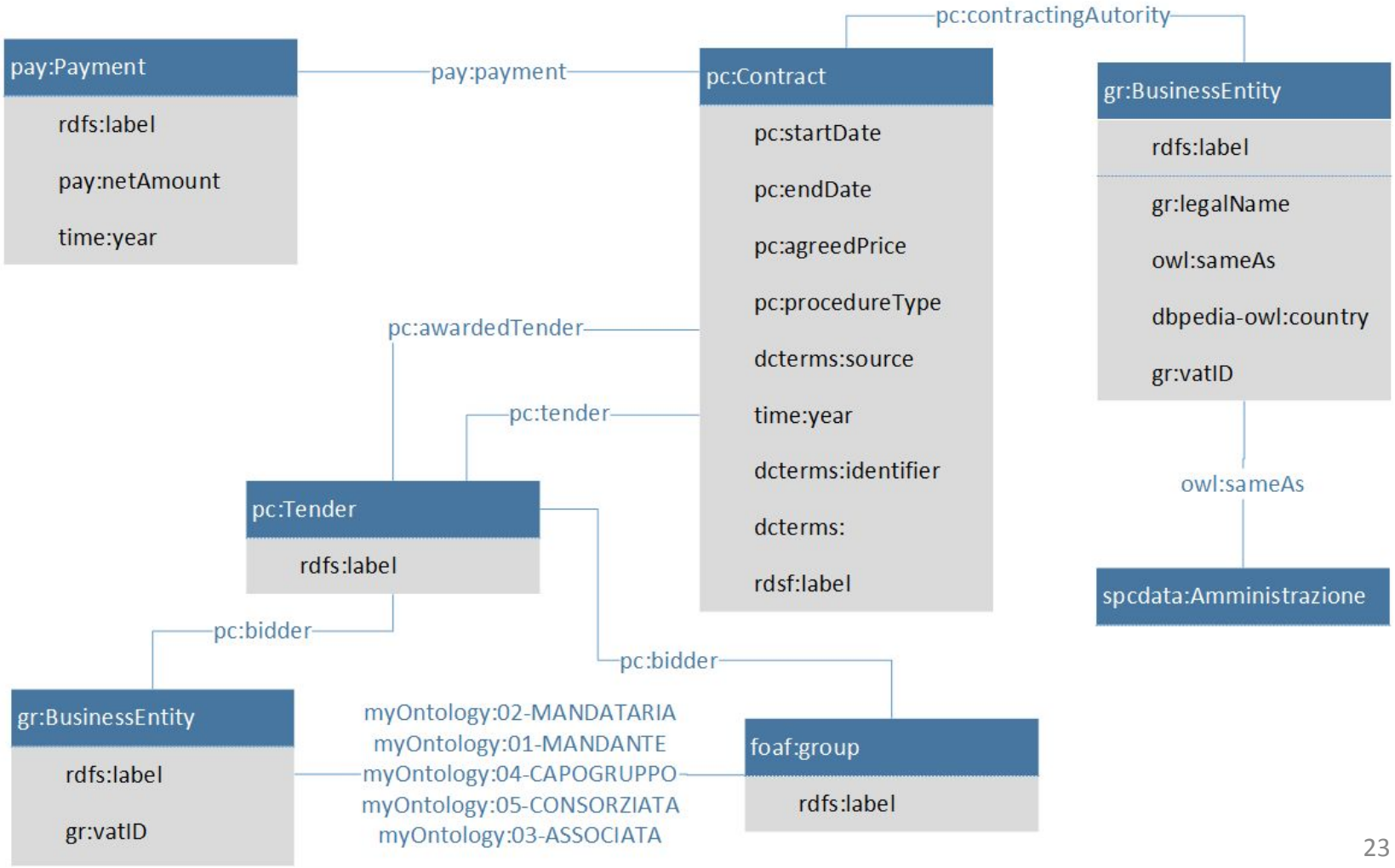
- Other types of inconsistencies **manifest themselves only after interlinking and merging data** contained in different sources
  - business entities with **more than one business name**
  - CIGs that identify **more than one contract**
  - **incoherent payments** among different versions of an ongoing contract

These issues represent **a significant barrier to achieve transparency**, because the results obtained by querying the dataset are misleading

Our Contribution


Web
Pipeline
Data
Italian Public Administrations websites (2)
dati.anticorruzione.it (1)
spcdata.digitpa.gov.it (8)
SPARQL endpoint (14)
ContrattiPubblici.org (15)
SPARQL
API
Downloader (3)
Cleaner (6)
Entity matcher (9)
RDF converter (10)
Web server (13)
API
XML
raw data (4)
MongoDB
download
metadata (5)
cleaned data
(7)
R D F
(11)
Lucene index
(12)

pc:	http://purl.org/procurement/public-contracts#
dcterms:	http://purl.org/dc/terms#
gr:	http://purl.org/goodrelations/v1#
dbpedia-owl:	http://dbpedia.org/ontology/
pay:	http://reference.data.gov.uk/def/payment#
time:	http://www.w3.org/2006/time#
org:	http://www.w3.org/ns/org#
foaf:	http://xmlns.com/foaf/0.1#


pay:Payment
rdfs:label
pay:netAmount
time:year
pay:payment
pc:Contract
pc:startDate
pc:endDate
pc:agreedPrice
pc:procedureType
dcterms:source
time:year
dcterms:identifier
dcterms:
rdsf:label
pc:contractingAutority
gr:BusinessEntity
rdfs:label
gr:legalName
owl:sameAs
dbpedia-owl:country
gr:vatID
owl:sameAs
spcdata:Amministrazione
pc:awardedTender
pc:tender
pc:Tender
rdfs:label
pc:bidder
pc:bidder
gr:BusinessEntity
rdfs:label
gr:vatID
myOntology:02-MANDATARIA
myOntology:01-MANDANTE
myOntology:04-CAPOGRUPPO
myOntology:05-CONSORZIATA
myOntology:03-ASSOCIATA
foaf:group
rdfs:label

Results

Table 2. Accuracy, completeness, and consistency degree in PP data

Field	Error	Occ. (%)	Solution
	Completeness		
Start date	missing	12.25	nothing
End date	missing	21.61	nothing
Agreed price	missing	0.06	nothing
Payment	missing	0.20	nothing
Procedure type	missing	0.11	nothing
Business Entity ID	missing	1,05	hash value
	Accuracy		
Identifier	syntactic errors	0.96	string cleaned
	semantic errors	5.83	hash value
Start date	semantic errors	1.36	nothing
End date	semantic errors	2.00	nothing
Agreed price	syntactic errors	0.94	string cleaned
	semantic errors	0.23	nothing
Payment	syntactic errors	0.76	string cleaned
	semantic errors	0.65	nothing
Procedure type	syntactic errors	2.81	optimal string match
Business Entity ID	semantic errors	1,08	hash value
	Consistency		
Start date	non standard format	5.63	uniformed to ISO 8601
End date	non standard format	5.20	uniformed to ISO 8601
Beneficiary	more than one beneficiary	1.78	nothing
Payment	payment without winner	4.30	nothing
	greater than awarded price	5.96	nothing

Table 2. Accuracy, completeness, and consistency degree in PP data

Field	Error	Occ. (%)	Solution
	Completeness		
Start date	missing	12.25	nothing
End date	missing	21.61	nothing
Agreed price	missing	0.06	nothing
Payment	missing	0.20	nothing
Procedure type	missing	0.11	nothing
Business Entity ID	missing	1,05	hash value
	Accuracy		
Identifier	syntactic errors	0.96	string cleaned
	semantic errors	5.83	hash value
Start date	semantic errors	1.36	nothing
End date	semantic errors	2.00	nothing
Agreed price	syntactic errors	0.94	string cleaned
	semantic errors	0.23	nothing
Payment	syntactic errors	0.76	string cleaned
	semantic errors	0.65	nothing
Procedure type	syntactic errors	2.81	optimal string match
Business Entity ID	semantic errors	1,08	hash value
	Consistency		
Start date	non standard format	5.63	uniformed to ISO 8601
End date	non standard format	5.20	uniformed to ISO 8601
Beneficiary	more than one beneficiary	1.78	nothing
Payment	payment without winner	4.30	nothing
	greater than awarded price	5.96	nothing

Table 3. Characteristics of Italian procurement information published as linked data

Dimension	Value
RDF triples	168,961,163
entities	22,436,784
contracts	5,783,968
public bodies	16,593
companies	652,121
links to external datasets	13,486

Discussion on Inconsistency Issues

- Business entities with **more than one business name**
- CIGs that identify **more than one contract**
- **Incoherent payments** among different versions of an ongoing contract

Business entities with more than one business name

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX gr: <http://purl.org/goodrelations/v1#>
SELECT (COUNT(DISTINCT ?be)) WHERE {
   {
      SELECT DISTINCT(?be) WHERE {
         ?be rdfs:label ?label .
        ?be a gr:BusinessEntity .
      }
    GROUP BY ?be HAVING (count(*)>1)
   }
}
```

Exploiting **VAT ID value**, we obtain a **unique business name** for contracting authorities, building links to the Italian public administration index of **SPCData** (http://spcdata.digitpa.gov.it:8899/sparql)

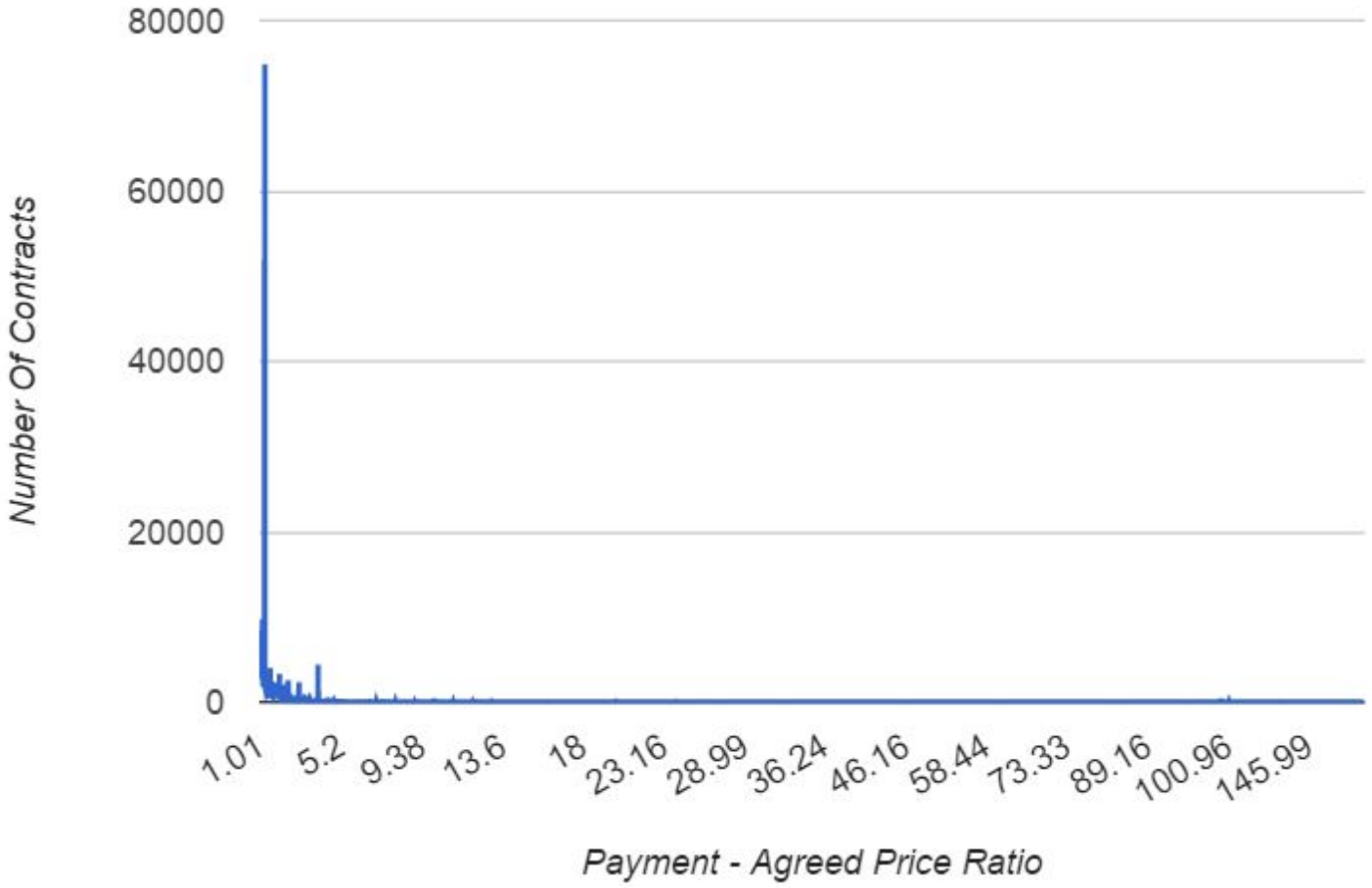
CIGs that identify more than one contract

```
PREFIX dcterms: <http://purl.org/dc/terms/>
PREFIX pc: <http://purl.org/procurement/public-contracts#>
SELECT (COUNT(DISTINCT ?contract)) WHERE {
   {
      SELECT DISTINCT(?contract) WHERE {
         ?contract dcterms:identifier ?CIG .
         ?contract a pc:Contract .
   }
   GROUP BY ?contract HAVING (count(*)>1)
   }
}
```

The solution to this issue is **generating a hash value**, avoiding ambiguity due to duplicate CIGs, to build contracts URIs. In this way, we separate different contracts, misidentified by the same CIG, in **different entities**

Incoherent payments among different versions of an ongoing contract

```
select ?result (COUNT(*) as ?n) where {
  {select ((ROUND(?pay_sum/?ap*100)/100) as ?result) where {
     ?c <http://purl.org/procurement/public-contracts#agreedPrice> ?ap .
     FILTER (?ap > 0).
     FILTER (?ap < 1000000000000).
     { select ?c (SUM(?pay) as ?pay_sum)
       where {
         ?c <http://reference.data.gov.uk/def/payment#payment> ?p .
         ?p <http://reference.data.gov.uk/def/payment#netAmount> ?pay .
         FILTER (?pay > 0).
         FILTER (?pay < 1000000000000). }group by ?c
     }}}} group by ?result order by ?result
```


80000
60000
40000
20000
0
Number Of Contracts
1.01
5.2
9.38
13.6
18
23.16
28.99
36.24
46.16
58.44
73.33
89.16
100.96
145.99
Payment - Agreed Price Ratio

Conclusions

- We presented an approach to tackle **fragmentation** of Italian procurement data and to **improve consistency** of such data based on **semantic technologies**

- Both these issues represent a **significant barrier to achieve a full transparency**, because user and robots that query the datasets risk to obtain partial, inconsistent, and misleading results

Future Works

- Developing **automatic tools** to detect and fix consistency problems among contracts published in different files

- Exploring ways to evaluate the **provenance of data**, in order to improve the data processing stage and improve data consistency

- Exploiting **advanced methods and dashboards** in order to monitor the consistency degree of the data

Thank you!

Mail: giuseppe.futia@polito.it

Twitter: giuseppe_futia

GitHub: https://github.com/synapta/public-contracts