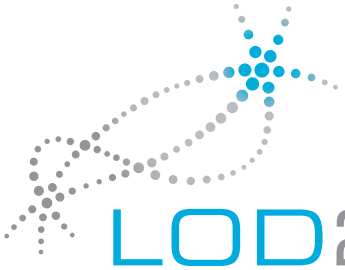


Collaborative Project

LOD2 - Creating Knowledge out of Interlinked Data

Project Number: 257943 **Start Date of Project:** 01/09/2010 **Duration:** 48 months

Deliverable 9a.3.1
Application of Data Analytics Methods on Linked Data in the Domain of PSC

Dissemination Level	Public
Due Date of Deliverable	Month 39, 31/11/2013
Actual Submission Date	Month 46, 08/06/2014
Work Package	WP9a, LOD2 for a Distributed Marketplace for Public Sector Contracts
Task	Task T9a.3
Type	Report
Approval Status	Accepted
Version	1.0
Number of Pages	39
Filename	D9a31.pdf

Abstract: Linked data on public contracts resulting from Task 9a.1 and Task 9a.2 has been submitted to different data analytics methods: from simple statistical aggregations through visual analytics to data mining in propositionalized representation and DL-based concept learning in structural representation.

Project funded by the European Commission within the Seventh Framework Programme (2007 - 2013)

History

Version	Date	Reason	Revised by
0.1	2014-03-06	Initial version	Vojtěch Svátek, Jindřich Mynarz
0.9	2014-05-08	Version for internal review	Vojtěch Svátek, Jindřich Mynarz, Krzysztof Węcel
0.95	2014-05-16	Reviewed	Heiko Paulheim
1.0	2014-06-8	Version for submission	Vojtěch Svátek, Jindřich Mynarz, Krzysztof Węcel, Sander van der Waal

Author List

Organization	Name	Contact Information
UEP	Vojtěch Svátek	svatek@vse.cz
UEP	Jindřich Mynarz	jindrich.mynarz@vse.cz
UEP	David Chudán	xchud01@vse.cz
UEP	Jakub Klímek	klimek@ksi.mff.cuni.cz
I2G	Łukasz Grzybowski	lukasz.grzybowski@i2g.pl
I2G	Mateusz Jarmużek	mateusz.jarmuzek@i2g.pl
I2G	Krzysztof Węcel	krzysztof.wecel@i2g.pl
ULEI	Lorenz Bühmann	buehmann@informatik.uni-leipzig.de
OKF	Sander van der Waal	sander.vanderwaal@okfn.org

Executive Summary

The deliverable describes the initial phase of applying different analytical techniques on public procurement linked data resulting from Task 9a.1. First, the state of the art is briefly described, spanning over the intersection of three topics: public procurement, linked data, and data analytics in general. Then the stakeholders are identified for which the overall task is relevant, and more specific analytic task types in the procurement domain are characterized and associated with the stakeholder types. The rest of the deliverable is then confined to the actual analytical studies undertaken. The three procurement data collections that have been addressed (Czech, US and Polish, each consisting of data from heterogeneous sources) are characterized and the process of obtaining the raw RDF data from the respective legacy sources is briefly described. Specific tools and toolboxes applied (or considered to be applied in the near future) are then characterized: visualization tools (Payola and B-Annot), propositional data miners (RapidMiner, WEKA, SAS Enterprise Miner and LISp-Miner) and relational data miners (DL-Learner and graph kernels). The targeted pre-processing of RDF data for the analytical tools and the actual (albeit, preliminary) analysis process/results, involving external LOD cloud data, are described. Finally, prospects for integration of the analytical functionality, based on experience thus gained, into the Public Contracts Filing Application (as subject of the follow-up deliverable) are outlined.

Abbreviations and Acronyms

CPV	Common Procurement Vocabulary
LOD	Linked Open Data
LD	Linked Data
OWL	Web Ontology Language
GR	GoodRelations (ontology)
PCO	Public Contracts Ontology
PSC	Public sector contracts
RDF	Resource Description Framework
URI	Uniform Resource Identifier

Table of Contents

List of Figures

List of Tables

1 State of the Art

The topic of the deliverable spans over three topics: linked data (as data source), data analytics (as type of data processing) and public procurement (as subject domain). Therefore, in the SoA overview, we briefly go through their intersections, pairwise as well as for all three.

1.1 Procurement Data and Analytics

A common EU-wide market for public procurement was implemented by the European Union in a directive in 2014.[1] In practice the regulation means that governments, municipalities and publicly owned companies such as utility companies are obliged under EU law to follow a strict tender process and to publicly publish all tenders above a certain threshold. The minimum thresholds under the EU directive varies from 130,000 EUR to 5,000,000 EUR depending on the type of contract.[2]

The database Tender Electronic Daily aggregates the tender information using data submitted from national procurement portals and publishes this on one single entry site. Tender Electronic Daily is maintained by the EU Publication Office, which therefore also holds the responsibility for implementing the technical solutions.[3] The data on Tender Electronic Daily includes several types of documents, such as Announcements of tenders, Contract Award notices, which declare the winner of a tender bid, as well as other types of announcements related to the tender process. While the data at Tender Electronic Daily has been free to access as aggregate via FTP server since January 1st 2014, it still does not provide bulk download of the data. Additionally accessing information requires a mandatory registration, which does not comply with the Open Definition.[4]

All tenders announced on Tender Electronic Daily includes a list of basic fields, which enable potential bidders as well as the public to navigate in the data. For example the Common Procurement Vocabulary – the CPV system – enables bidders to understand which tenders are available within their own sector.[5]

For civil society organizations and journalists Contract Award notices have gained particular attention as these includes information about the winner of the tender (incl. company name, company address, contract amount). Researchers have however raised criticism of the data collection conducted by the EU Publications Office, mainly focusing on missing data within Contract Award notices, for example missing dates and missing amounts.[6]

Citing from a blogpost by Anders Pedersen:[7] "Procurement data tended, for good reasons, to enjoy much less attention from journalists than for instance spending data. A major reason is the fact that contract information tend to be more about text than spreadsheets, but also because access to aggregate data has been limited. However with access to more than 100,000 public sector contracts annually from the European procurement register originating from tiny municipalities to large government agencies, there are good reasons to explore if contracts can help fill out the blanks." According to the same source, procurement data are "a solid source for single stories" told by journalists as well as transparency NGOs. On the other hand, aggregate analysis known from public spending data is not common on procurement data. It mostly focuses on a change of a

[1] http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:2004L0018:20120101:EN:PDF
[2] http://europa.eu/legislation_summaries/internal_market/businesses/public_procurement/l22009_en.htm
[3] http://ec.europa.eu/internal_market/publicprocurement/index_en.htm
[4] http://opendefinition.org/
[5] http://ec.europa.eu/internal_market/publicprocurement/rules/current/index_en.htm#maincontentSec6
[6] http://chaire-eppp.org/en/node/602
[7] http://community.openspending.org/2013/04/procurement-hack-day/

single measure (such as the overall count or financial amount) over time, as in the Czech study by zIndex,[8], examining the effect of a new legislation on the financial amount of public contracts (globally as well as per sector, plus in comparison with public spending).

The most recent analytical activity for TED, the Brussels Hack Day on EU Procurement data[9] (May 2014) already examined structured CSV data extracted from the TED XML files within the OpenTED initiative.[10]. The discussion however focused on the quality and completeness of the data rather than on proper application of analytical methods. It seems that aggregate analysis is still prevented by the fact that even in this comprehensive data source, "many pieces of essential information are missing – including many contract values and supplier names", and that the data is "very messy, particularly when it comes to clearly identifying the public body and economic operator involved in a contract".[11] Presumably, linked data technology could thus serve as major enabler here.

1.2 Use of Linked Data for Public Procurement

The potential utility of transforming public procurement data to the linked data format is obvious. First, it could then be effectively integrated with other linked government data, as well as encyclopedic and geographic data, leading to mutual enrichment. Second, although the original sources of procurement data are typically in tabular/hierarchical structure, the domain is closely connected with that of business network analysis (especially when it comes to transparency investigations), where the data are naturally graph-shaped; the integration overhead may then be lowered for RDF as graph-oriented language. Third, lightweight matchmaking with business offers now increasingly published on web pages in RDF (using vocabularies such as Schema.org or GoodRelations) is enabled by the transformation.

One of the first systematic attempts to transform procurement data to RDF, as early as 2010, was the LOTED project,[12]. Its first phase aimed to extract data on tenders in European Union coming from the Tenders Electronic Daily (TED)[13] portal via RSS feeds. A dedicated vocabulary was designed for this purpose[14] and a custom scraper was developed. Simple statistical analyses have been applied on the data as published over a SPARQL endpoint, enriched with relevant linked data from DBpedia and Geonames. The second phase of this project focused on placing the public contracts data into a legal context, through a fully-blown ontology [7].

A second notable attempt to apply linked data principles in the procurement domain was the MOLDEAS (Methods on Linked Data for E-procurement Applying Semantics) project [3]. Its overall architecture [26] took into account diverse sources, such as TED, BOE – the official bulletin of Spanish government, or BOPA – the official bulletin of the local government of the Spanish region of Asturias. The data coming from disparate sources were combined using linked data technologies to allow for application of semantic methods, such as SPARQL query expansion and query performance optimization [25] or spreading activation techniques used for information retrieval in graph data [4]. Controlled vocabularies, such as public sector code lists, product classifications or thesauri such as EUROVOC[15], were transformed to RDF, interlinked with external datasets, and used for indexing of public contracts data to enable intelligent services, such as complex subject-oriented queries. Query expansion was carried out: correlated product classification codes were included and numerical

[8] http://zindex.cz/data/2014-02-20-studie-objem-zakazek.pdf (in Czech)
[9] http://www.eventbrite.co.uk/e/hack-day-on-eu-procurement-data-tickets-6157668753
[10] http://ted.openspending.org/
[11] http://pudo.org/blog/2014/05/15/opented.html.
[12] http://loted.eu/
[13] http://ted.europa.eu/
[14] http://loted.eu/ontology
[15] http://eurovoc.europa.eu/

values (such as financial value or contract duration) were fuzzified so as to also cover values sufficiently close to the initial requirement (provided other parameters are favorable).

1.3 Linked Data Analytics and Mining

The mainstream method of accessing linked data in RDF is by means of SPARQL queries. However, individual queries do not allow to explore a large space of possible regularities. A deeper insight into data can be obtained via *graphical interfaces* that dynamically translate user actions into queries. Another type of linked data analytics is by automated *data mining* methods. This option has, however, only been investigated to a small extent, mostly in connection with RapidMiner, probably the most popular academic data mining tool.

Kiefer et al. [14] proposed SPARQL-ML, an approach to data mining on semantic web data focused on statistical relational learning and SPARQL. Similar direction has been followed by the RMonto tool [22], an upper layer for RapidMiner tool. It allows to apply ontologies as background knowledge for several mining tasks, possibly combining relational and propositional subtasks. Another implementation of RDF data pre-processing for RapidMiner is by Khan [12]. Finally, Paulheim & Fürnkranz [21] suggested an automated method for data enrichment from Linked Data, pipelining entity recognition, feature generation and feature selection. The original FeGeLOD framework has been later reengineered in the form of a plug-in for RapidMiner [20]. While the former approaches regard RDF data as relational data, Paulheim's approach attempts to bring linked data within the reach of propositional mining algorithms. Given a dataset that is already in propositional format, entities occurring in it are searched for on the LOD cloud. Data about these entities, collected from LOD resources (for example, encyclopaedic ones such as DBpedia and Geonames) are then transformed to additional features in the propositional table. If the number of features grows extensively, a feature selection step is then also needed.

Another proposition of Kiefer et al. [13] is iSPARQL – extension of traditional SPARQL with similarity measures that provide support to finding alike entities in Semantic Web knowledge bases. Proposed framework explores mainly three SPARQL-based aspects, such as semantic data integration, ontology mapping, and semantic web service matchmaking. It tests possibility of integration customized similarity functions into popular RDF query language. It makes use of Java library called SimPack,[16] which supplies measures from various categories: feature vectors, strings, sets, sequences, trees, graphs and information theory. Definition of the proximity of objects is crucial part of various analysis methods, especially for data mining clustering task, thus proposed framework seems to be interesting.

Nonetheless, it is possible to apply different solution relative to early mentioned approaches. Kavitha et al. [11] presented three basic trends of graph mining categorized based on the approach used to search frequent subgraphs in large graph data set:

- incomplete beam search greedy method – use heuristics to evaluate optimal solution construed in stages
- inductive logic programming – utilize logic for data representation and search
- mathematical graph theory – explore a complete set of subgraphs mainly using a support or a frequency measure.

Graph representation has great expressiveness at the expense of performance from the complexity of representation, access and processing points of view. An alternative way can be transformation from graph representation data set to a single-table form (called propositionalisation process) which allows using classic data mining techniques. Main problem of this case is kind of information loss due to simplification of data

[16] https://files.ifi.uzh.ch/ddis/oldweb/ddis/research/simpack/

representation as a consequence of limited table form expression. It is noticeable particularly when we are dealing with sequential, temporal, spatial, structural, multidimensional or hierarchical data.

Mining frequent associations directly from *RDF statements* has been investigated by Abedjan & Naumann [1], who proposed the notion of mining configuration, specifying the position (in the triple) of the entities to be associated (so called target) and of the entities whose frequencies establish the association (so called context); for example, associations (unidirectional similarities) between subjects can be found based on frequently co-occurring predicates.

1.4 Procurement linked data analytics

The pioneering effort in analytics over procurement linked data was in the mentioned LOTED project. The developed application (no more online) allowed the users to specify a tender profile and compare tenders according to their various dimensions. The user could compare data across the EU member countries submitting tenders to TED. Given the multidimensional character of the data, one could then try to find correlations between tenders' profiles and various other features, including location or political party.

2 Stakeholders in Procurement Linked Data Analytics

Although the task of large-scale analytics over linked data is rather new, it is likely to inherit its stakeholders from its constituent areas. We identify at least five distinct groups.

NGOs/Journalists. This is probably the largest group that primarily addresses open procurement data with analytical targets. While a large part of the 'watchdog' effort, aiming to reveal corruption and clientelism in public sector, deals with unstructured data that need to be processed by humans, structured data repositories can also provide them important insights.

Official government bodies. In many countries, there are both specific supervisory bodies that address the issues of transparency and fair competition and statistical offices that merely collect data as part of aggregated information on the national economy.

Bidders. In the context of bidding, analytics is an activity closely tied with matchmaking. The organizations that intend to submit a tender may naturally be interested in continuously finding the best opportunities in their sector. Analysing the previous successful and unsuccessful tenders may be then helpful. In long term, the companies may wish to actively plan their bidding strategies based on procurement market trends, as revealed by automated analysis.

Contracting authorities. Similarly, on the other side of the market, the contracting authorities as future buyers may want to understand the supply side in order to know how to formulate the contract conditions, in view of successful matchmaking. Good progress of a future contract may be derived from previous experience with certain bidders. An additional goal may be to attract an adequate number of bidders; excessively many bidders bring large overheads to the awarding process, while too low a number may reduce competition (and, under some circumstances, even lead to contract cancelling by a supervisory body, due to an anti-monopoly action).

Researchers on analytical methods. Even if this stakeholder group is not specifically tied with the procurement domain, they may find it favorable even to them in terms of demonstrating a practical impact of their research to applied research funding agencies. Besides, linked data is a novel type of data for the data mining community. It breaks down many traditional assumptions on source data and thus represents a number of challenges:

- While the individual published datasets typically follow a relatively regular, relational-like (or hierarchical, in the case of taxonomic classification) structure, the presence of semantic links among them makes the resulting 'hyper-dataset' akin to general graph datasets. On the other hand, compared to graphs such as social networks, there is a larger variety of link types in the graph.
- The datasets have been published for entirely different purposes, such as statistical data publishing based on legal commitment of government bodies vs. publishing of encyclopedic data by internet volunteers vs. data sharing within a researcher community. This introduces further data modeling heterogeneity and uneven degree of completeness and reliability.
- The amount and diversity of resources as well as their link sets is steadily growing, which allows for inclusion of new linked datasets into the mining dataset nearly on the fly, at the same time, however, making the feature selection problem extremely hard.

Linked data in general, and procurement linked data in particular, is thus likely to become one of important meeting points of data mining and other analytical tools, especially those capable of consuming 'non-rectangular' (relational, or graph-based) data.

3 Tasks Considered in the Study

3.1 Descriptive Tasks

3.1.1 Simple Aggregation of Past Contract Data

Basic SPARQL-based aggregation of past contract data provides simple statistics helping to get an overview of a particular dataset. There are several subtasks typically carried out in this context:

- application of SPARQL queries to retrieve data
- application of SPARQL queries to calculate statistics
- integration of SPARQL queries results with visual components.

Statistics can include number of contracting authorities, contractors, contracts, lots, or addresses. Value of contracts can be calculated as sum or average per authority, contractor, region, kind of delivery, classification of goods etc. Charts can be generated for presentation of these statistics split by various dimensions (e.g. bar charts) or showing the evolution (e.g. line charts, timeline). Geographical dimension is best presented on maps – detailed data can be shown as points on the map, e.g. pointers with shaded tooltips on OpenStreetMap; for aggregated data we need to prepare chorophlet maps. Sample application using the above methods is presented in section 6.1.

3.1.2 Clustering: Looking for Similar Contracts

Clustering consists of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters. There is no agreed definition of clustering as the notion of cluster depends on the goal of research and thus cannot be universally defined. As a consequence, there are many clustering algorithms such as connectivity based clustering (hierarchical clustering), centroid-based clustering (k-means clustering), distribution-based clustering, density-based clustering [8].

As a primary object for clustering we consider contract notices. At this moment we do not specify whether it is strict partitioning (clusters are disjoint) or overlapping clustering; whether there may be outliers or clustering is exhaustive. Also, there can be different interpretation of clusters depending on the person looking at it. Here, hierarchical clustering can be a good solution, i.e. number and details of clusters can be parametrized. For this, appropriate distance function has to be defined.

The main beneficiary interested in this challenge are the bidders. The business case for them is rather obvious. When they look for new opportunities it is possible to compare new notices to the contracts they have already realised or they may identify contracts from the past that would be most suitable for them. Currently they have the possibility to monitor new contracts based on basic criteria but the "search language" is not very expressive. Also, the most popular criterion – CPV – may be misleading as some contracts may be incorrectly classified.

Clustering also offers a business case for contracting authorities: from time to time they have to find past contracts similar to their current need, so that they can prepare their own notice. In the case of the big companies the processes are relatively stable, and not many new types of contracts have to be announced. For example, a university can look for similar contracts by other universities. Here the use of CPV is not efficient.[17]

[17] Source: survey with dedicated departments at a university.

When future contracts are concerned, looking for similar contracts can help to identify aggregated demand opportunities.

3.1.3 Outlier Detection

After the clustering is done, we can find contracts that do not fit into any of the clusters. Such contracts are exceptional when taking into account many criteria, e.g. high number of bidders, unusual CPV code, outstanding price. The identification of outlier contracts heavily depends on the set of attributes considered.

Unsupervised detection of outliers makes an interesting case for analyses. The contracts identified in this analysis should be further investigated as they make an entry point for suspicious behaviour or even fraud.

3.1.4 Association Mining

By applying association mining techniques one can discover ties between contracting authorities, bidders and offers. Association-related techniques are of topmost interest for supervisory bodies. They make it possible to discover any abnormalities on the market. There are several crucial relation types; their judgement depends on objects of analysis. The following relations are the most interesting:

- between contractors and products: stability of the offer; the tighter the relationship, the more reliable the contractor is
- between contractors and authorities: it signals the need to check for corruption.

There are many algorithms for mining association rules, from Agrawal's initial idea [2] to more recent algorithms executed directly in SPARQL [16]. A specific family of association rule mining methods is based on observational calculi derived from the GUHA method [23], implemented in full in the LISp-Miner system.[18]

The overall public contract data allows deeper economic analyses, e.g., the depth of the market: the market is shallow when some products or services are offered by a small number of contractors.

3.2 Predictive Tasks

The techniques presented so far are classified as unsupervised learning methods. Predictive models, in turn, require supervised learning, i.e. labelled data has to be provided. This can be achieved based on historical contract notices:

- closed notices, i.e. with contract award notice, may constitute the learning dataset
- open notices, i.e. with closing date in the future, may constitute the scoring dataset.

To be more precise, closing date is suitable criterion for dividing the data between the learning vs. scoring dataset when a feature established at the time of notice closing is to be predicted, such as the number of bidders mentioned below. For contract winner prediction, in turn, the award date would typically be critical. Finally, there can be also atemporal classification tasks, such as the classification of contracts as multicontract; this is an expertise-guided classification without ground truth determined in an unambiguous way.

[18] http://lispminer.vse.cz

3.2.1 Prediction of Number of Bidders

Finding a good business case for prediction is not straightforward as public procurement data is mostly qualitative. There is, however, at least one variable that can be of primary interest for contracting authorities – number of bidders. The authorities that 'play fair' are interested in awarding the contract as quickly as possible. They want to avoid long processes due to protests and legal issues handling. Organizing contracts many times (e.g., no offers submitted) is also not an option as it all causes wasting time and costs money. Theoretically, the higher number of tenders, the better, although excessive number of tenders may lead to unnecessary overhead for the procurement authority.

Supervisory bodies may be interested in contracts with just one bidder. Overspecified contracts may reveal an intention to overcome the public procurement law when a winner is determined before a contract notice is filed.

3.2.2 Multi-Contract Prediction

A multi-contract is a contract that (often, 'suspiciously') unifies two or more unrelated commodities. Multi-contract prediction and inspection will be useful for detection of fraud, waste of public resources and corruption. By law, dividing contract into smaller contracts or parts and understating the estimated value of contract is forbidden. Contracting authorities may, in this way, understate the contract's value to keep it below the threshold requiring them more detailed publishing.

3.2.3 Successful Tender Prediction

This is definitely the most difficult prediction tasks but at the same time gaining the most interest from business perspective. Successful tender prediction cannot be done without additional weakening conditions or apposition of the case.

The main problem is access to data. In most cases it is not possible to obtain 'negative' examples, i.e. we know who is the winner but we do not know who else bided. In fact, the prediction should not only base on the winning organisations but should also consider co-tenders as well.

To certain extent it is possible to identify a group of potential candidates by analysis and interpretation of historical data. We can build a profile of potential contractor and match it to the contract notice.

One scenario to be used by contracting authority is to identify potential bidders. This information can be valuable when inviting to closed procedure. The other scenario to be used by bidders is to get an overview of competitors, and based on that also estimate the price level by looking into historical data. Even though the prediction is not very reliable, it gives some insights into the expected success. A company can then decide to get involved into the process and make a bid.

4 Data Acquisition and RDFization

Although procurement data is published, in a certain form, in most countries in the world, in the LOD2 WP9a we focused on three groups of sources:

1. US and UK procurement data portals, as this is the countries where the open government publishing campaign started first and therefore even the procurement data sources are likely to be found sufficiently rich and well curated.
2. The Czech and Polish procurement data portals; the lead partners of WP9a are based in these two countries, and therefore have with both good contacts to the national publishing agencies, knowledge of the local regulations, and fluency in the languages in which the unstructured part of the data is written.
3. The European TED portal, which contains data from a number of countries (thus allowing for cross-country comparisons, as shown in [29]), although only a subset of these (typically for contracts above a certain price level).

While TED had been the prime target in the initial phase of the project, the ETL activities for this resources were later suspended when the future availability of full data in XML (rather than mere HTML) was announced. The processing of TED data was thus only resumed in Spring 2014 based on XML dumps, which are more reliable than data obtained via information extraction from semi-structured text embedded in HTML. Furthermore, the UK data from the ContractsFinder portal[19] lack one of our major points of interest: the number of tenders. The results presented in this deliverable thus mainly rely on the US, Czech and Polish procurement data.

4.1 U.S. Data

The main purpose of collecting and transforming U.S. procurement data to RDF was to acquire a dataset suitable for the tasks defined for the Linked Data Mining Challenge (LDMC). The U.S. dataset was used both for the 2013 [28] and 2014[20] editions of LDMC. The reasons why we reached out for data not coming from the EU are due to its rich content and availability in machine readable formats. The data is sufficiently rich to support the requirements of LDMC as it contains numbers of bidders who tender for public contracts, which was the variable sought in LDMC's predictive tasks. Moreover, the data is available in machine readable formats and with unrestricted terms of use given that it is in public domain.

Similarly to Czech data, the U.S. dataset was created by combining data from two principal sources, which provide complementary kinds of data. The two sources in question are USASpending.gov[21] and Federal Business Opportunities (FBO).[22] USASpending.gov offers a database of government expenditures, including awarded public contracts, for which it records e.g., the aforementioned numbers of bidders. On the other hand, FBO publishes public notices for ongoing calls for tenders. Once public notice's deadline for tender submission passes, final number of bidders should be published along with other information about contract award in USASpending.gov. Unfortunately, these two sources do not publish enough data about public contracts to pair the equivalent instances reliably. While the same contract identifiers are used in some cases, most of the published contracts lacks identifying information necessary for deduplication. Combination of data from the

[19] http://contractsfinder.businesslink.gov.uk
[20] http://knowalod2014.informatik.uni-mannheim.de/en/linked-data-mining-challenge/
[21] http://usaspending.gov/
[22] https://www.fbo.gov/

two sources thus yields only a small subset of public contracts that could be merged provided they are equipped with strong identifiers, such as URIs.

USASpending.gov provides data downloads in several structured data formats, including CSV, TSV, XML and Atom. We used the CSV dumps, which we converted to RDF using SPARQL mapping[23] executed by tarql.[24] Data dump from FBO is available in XML as part of the Data.gov initiative.[25] To convert the data to RDF we created an XSLT stylesheet that outputs RDF/XML.[26] As additional dataset using in both USASpending.gov and FBO, we converted the FAR Product and Service Codes[27] to RDF using LODRefine.[28]

Data resulting from transformation to RDF was interlinked both internally and with external datasets. Internal linking was done in order to fuse equivalent instances of public contracts and business entities (both contracting authorities and bidders). Deduplication was performed using data processing unit for UnifiedViews that wraps Silk link discovery framework.[29] The output links were merged using data fusion component of UnifiedViews.[30] Links to external resources were created either by using code-based URI templates in transformation to RDF or by instance matching based on converted data. The use of codes as strong identifiers enabled automatic generation of links to FAR codes and North American Industry Classification System 2012,[31] two controlled vocabularies used to express objects and kinds of public contracts. Instance matching was applied to discover links to DBpedia and OpenCorporates.[32] Links to DBpedia were created for populated places referred to from postal addresses in the U.S. procurement dataset. In this case, the employed Silk linkage rule was based on comparison of normalized ZIP codes and pre-filtering possible matches by transitively-expanded Wikipedia category for populated places in the U.S. OpenCorporates was used as target for linking bidding companies. The task was carried out using batch reconciliation API of OpenCorporates via interface in LODRefine. Links were established based on pre-filtering by jurisdiction and fuzzy matching on normalized legal name, with which company is registered in respective jurisdiction. In all cases of instance matching samples of resulting links were verified by manual scrutiny, in order to estimate linking accuracy.

U.S. dataset characteristics:

- Number of triples: 18.9 M
- Number of notices: N/A
- Number of contracts: 186,364
- Number of contracting authorities: 648
- Number of contractors: 48,855
- Number of business entities: 58,781
- Number of unique NAICS codes: 1,269

[23] https://github.com/opendatacz/USASpending2RDF
[24] https://github.com/cygri/tarql
[25] ftp://ftp.fbo.gov/datagov/
[26] https://github.com/opendatacz/FBO2RDF
[27] http://www.acquisition.gov/
[28] http://code.zemanta.com/sparkica/
[29] http://wifo5-03.informatik.uni-mannheim.de/bizer/silk/
[30] Developed previously for ODCleanStore, the predecessor of UnifiedViews [19].
[31] http://www.census.gov/eos/www/naics/index.html
[32] https://opencorporates.com/

4.2 Czech Data

Public procurement data from the Czech Republic that we processed comes from two main sources. Since the start of our involvement with the LOD2 project we worked with data from the central Czech public procurement register.[33] Later on, as revised public procurement legislation came into force in 2013, an additional data source appeared in the form of profile feeds that contracting authorities were obliged to start publishing.

Until the autumn of 2013 the data extraction from the central Czech public procurement register involved scraping HTML pages and distilling structured data out of them via CSS selectors, regular expressions and other means of processing semi-structured text. Late of 2013 the company maintaining the register[34] opened up an XML-based SOAP API to both read and write data from the register. We acquired access to the API and ceased to scrape the HTML pages of the register, as the API promised a more reliable way of obtaining data. At the turn of the year we developed an extractor data processing unit[35] for the UnifiedViews ETL framework,[36] which is capable of incremental extraction of data from the register using its SOAP API. During the time we discussed the possibility of publishing raw open data in bulk with the company running the register. As a result of these discussions we were provided with XML dump of historical data from the register to be used for research purposes. Combining the historical data dump with the access to current data via the SOAP API we were able to reconstruct the complete dataset of public contracts from the registry converted to RDF.

The second source of Czech public procurement data that we processed was a set of profile feeds of individual contracting authorities. As per the amendments in the Czech public procurement law, public sector bodies involved in public procurement are required to publish their own XML feed of data about public contracts they issue, including both public notices and award information. The set of public contracts that are published on profile feeds is a superset of what is available via the central Czech public procurement registry because the feeds also cover some lower price public contracts, which are not required to be published in the central register. Content of these feeds mostly mirrors the content of the central register, although for individual public contracts it is less comprehensive. Even though the data from the register is richer and more descriptive, the profile feeds contain information about unsuccessful tenders, which is missing from the register that only reveal information about winning tenders. We deem having data about both successful and unsuccessful tenders as vital in several analytical tasks over public procurement data, which is one of the reasons why we have invested effort into acquiring the data from feeds of contracting authorities. Since early autumn 2013 we have been scraping an HTML list of URLs of profile feeds and periodically convert each feed's XML into RDF using an ETL pipeline developed using the UnifiedViews framework.

Public procurement data from both these sources is modelled in RDF using the Public Contracts Ontology [15, p. 21], using an extension module developed for several aspects specific to the Czech setting. By using code-based URIs the data is linked to several external datasets. Company identifiers connect it to the Czech business register that we also periodically convert to RDF. Common Procurement Vocabulary (CPV) codes link it to the RDF version of CPV that we produced.[37]

Czech dataset characteristics:

- Number of triples: 11 M
- Number of notices: 113,663
- Number of contracts: 181,258

[33] http://vestnikverejnychzakazek.cz/
[34] http://www.ness.com/
[35] https://github.com/opendatacz/VVZ_extractor
[36] https://github.com/UnifiedViews/Core
[37] Earlier, we used RDF version of CPV developed by WESO Oviedo research group. However, it is no longer maintained, so we converted CPV to RDF ourselves and fixed several bugs introduced in the vocabulary's hierarchy in the process.

- Number of contracting authorities: 116,808
- Number of contractors: 108,096
- Number of business entities: 227,170
- Number of unique CPV codes: 6,232

4.3 Polish Data

Polish public procurement data is published by The Public Procurement Office (Urząd Zamówień Publicznych – UZP[38]) in the Public Procurement Bulletin (Biuletyn Zamówień Publicznych – BZP[39]). By Polish Public Contracts' law, data should be open. There are several means to access data: browsing the BZP portal (HTML), subscription mechanism with some restricted number of criteria (e-mail), or direct data download (XML files).

We decided to follow the third way to access data, i.e. download of raw data in XML files[40]. It was a reasonable move by The Office to publish public contracts notices as daily XML dumps – by offloading files to external server they avoid overload of main server by people trying to scrape data from HTML pages. It was necessary to transform this collection to graph representation for Linked data analysis. We recognized several problems related to transformation task. Some XML files contained illegal unicode characters, which had to be removed. There were also some unescaped ampersand signs, causing confusion with XML entities. These serialization problems have been solved by dedicated script in Python that finds all incorrect files and fixes them. Structure of XML was not designed with easy processing in mind, so mapping task was not straightforward.

Throughout the history of contract notices publication there were changes in the data model. Some of boolean character fields are coded in two ways – once by symbols *T* and *F*, another time by numbers *0* and *1*. However, the most serious problem is lack of normalisation of text fields filled by people. Some of them are totally useless with regard to the possibility of generalisation, classification and comparing entities. Certain of key data such as address and corporate name was repeatedly mentioned in various ways, once using abbreviation, once by full name, another time in incorrect form. It is not always obvious how to automate findings and repair such cases. One of possible solutions for some problems is employment of various dictionaries and similarity measures for unification of data notation.

Data process was composed of five parts:

- download XML files
- unpack files
- clean – remove illegal characters from XML files
- mapping process, and
- upload RDF to Virtuoso.

XML entities were mapped to Public Contract Ontology (PCO) and the other proper vocabularies – the process was mostly similar or the same as for Czech data. Polish data was additionally enriched with geographical vocabularies (NUTS and Polish TERYT), thus allowing basic aggregations across this dimension. Addresses available in the dataset were geocoded thus allowing precise location of contracting authorities and contractors on the map. CPV codes were used to standardize description of the subjects of procurement contracts. Business entities were identified using identifiers assigned by tax office (NIP) and statistical office (REGION).

[38] http://uzp.gov.pl
[39] http://uzp.gov.pl/BZP/
[40] ftp://ftp.uzp.gov.pl/bzp/xml/

Specificity of Polish public contracts is in the scope of data. There are nine types of notices: contract notice (ZP-400), simplified contract covered by the dynamic system – DPS (ZP-401), construction contract notice (ZP-402), contract award notice (ZP-403), notice about competition (ZP-404), announcement of the results of the competition (ZP-405), change of the announcement notice (ZP-406), service concession notice (ZP-407), public administration (ZP-408) and notice related to military units (ZP-409). There are 17 common attributes, present in all notices. ZP-400 is special with regard to number of distinctive attributes – there are 147 of them. While many of them are just binary flags, there are also some free text attributes that need to be analysed using text mining techniques. For this reason we created a Polish module extending the PCO with specific properties. It allows to represent all information from XML files in RDF format opening opportunities for future applications.

For the conversion process we used a framework that proved to be extremely powerful, flexible and relatively easy to use – Tripliser[41]. It is a Java library and command-line tool for creating triple graphs from XML. It is particularly suitable for messy, bulky, and volatile data. Designed as an alternative to XSLT conversion it provides easy-to-read mapping format, robustness (error or partial failure tolerance), and efficiency. Some examples of flexibility utilised in our extraction procedures include wildcard used to match unknown tags, built-in functions (e.g., `fn:encode-for-uri`, `fn:year-from-date`), own extension function (e.g., `myfn:encode-executor`), advanced XPath expressions (e.g., `parent::*/parent::*/` to navigate back in hierarchy). We also leveraged the possibility to provide own extension functions. For example, we had to generate identifiers based on several attributes for addresses (`encode-address`) and contractors (`encode-executor`). In order to avoid collisions, hashes based on several attributes were calculated. On one hand, we thus avoided multiple storage and the same data (blank nodes do not merge in Virtuoso), and on the other hand, merging of variants was possible by homogenizing the spelling, e.g., removal of dots, spaces, dashes, Polish letters and common errors.

Even though the data was contained in XML, several issues were encountered. First of all, XML was not designed with easy processing in mind. One of the problems is a possibly unlimited list of tags, e.g., for `wykonawcy` (executors) consecutive numbers are used: `wykonawca_0`, `wykonawca_1`, `wykonawca_2` and so on. Some XML files contained illegal unicode characters[42]. They have to be removed, otherwise XML parser does not work. Similarly, there were some unescaped ampersand signs, causing confusion with XML entities.

Polish dataset characteristics for 2013:

- Number of triples: 28.8 M
- Number of notices: 413,382
- Number of contracts: 922,038
- Number of contracting authorities: 17,648
- Number of contractors: 177,136
- Number of business entities: 194,784
- Number of unique CPV codes: 18,156

[41] https://github.com/daverog/Tripliser
[42] For example: /u0x1, /u0x2, /u0x3, /u0x4, /u0x5, /u0x6, /u0x7, /u0x15, /u0x19, /u0x1a

5 Data Pre-Processing for Analytics

5.1 Pre-Processing of Czech Data

As the selected analytical tools pose different requirements on data, we pursued two different directions in pre-processing the Czech public procurement data. Most of the analytical tools are not suited to work with RDF natively and thus transformations to non-RDF data formats needed to take place. On the other hand, analytical tools based on description logic, such as DL-Learner,[43] require data with rich ontological description, so then T-Box enrichment was necessary.

The basic data format that many of the analytical tools work with is propositional table. For example, such tabular data may be encoded in CSV format. A convenient approach for converting RDF into tabular form is SPARQL SELECT query. Given the functionality included in the SPARQL 1.1 Query specification,[44] it is also an expressive approach for pre-processing data, such as computing aggregated values. Moreover, most up-to-date RDF stores implement features of the SPARQL 1.1 Protocol,[45] which enables to retrieve SPARQL SELECT, results directly in CSV or TSV.[46]

We created several SPARQL SELECT queries spanning multiple datasets covering data relevant to Czech public procurement, such as the contracting authorities' profile feeds or the business register. Each of the queries focused on different aspect of the domain in question, such as aggregations per CPV code and geographic area or queries taking into account the age of companies involved in public procurement. The results of these queries were exported in CSV, so that they can be readily imported into data mining tools, such as EasyMiner[47] or RapidMiner[48].

The other direction of pre-processing was directly spawned by specific requirements of DL-Learner. This tool yields a lot of leverage from OWL axioms in ontologies or RDF vocabularies with which the provided data is described. However, in the case of Public Contracts Ontology (PCO), which is designed rather as a light-weight linked data vocabulary, some of the axioms required by DL-Learner[49] were missing, such as the distinction between datatype and object properties. Therefore we equipped a DL-Learner-specific version of PCO with these ontological constructs.

5.2 Pre-Processing of U.S. Data

The U.S. dataset was primarily meant for the participants of the Linked Data Mining Challenge. The three groups involved in the 2013 and 2014 editions proceeded as follows to be able to apply the data mining tools (details are in the cited papers).

University of Darmstadt (2013 participant, [18]) used the following heuristics when converting the RDF data to attributes of *Contract* objects:

[43] http://dl-learner.org/Projects/DLLearner
[44] http://www.w3.org/TR/sparql11-query
[45] http://www.w3.org/TR/sparql11-protocol
[46] Tab-separated values
[47] http://easyminer.eu
[48] http://rapidminer.com
[49] In fact, they were required by the OWL API, on which DL-Learner is based.

- numerical attributes were constructed from data properties pointing to literals of type *xsd:int*; if most occurrences were of this type, even the remaining ones were converted to numerical values if possible (such as string values consisting of qualifiers such as 'nearly', followed by a numeral)
- nominal attributes were constructed from data properties having, for all instances, a literal value shorter than 20 characters and the number of distinct values less than 30% of the number of all values
- the month value was extracted from each date-valued data property as a specific numerical attribute
- in the case of multiple objects for the same property, priority was given to the first occurring one;[50] however, in presence of language tags in literals, priority was given to those in English (*en tag*)
- the remaining attributes were left as string ones
- the rules above were also applied recursively to all data properties connected to contract instances via object properties (up to four hops); the property chain involved became a specific attribute; however, *owl:sameAs* property was not used in this sense as it leads to rapid increase of multi-valued attributes
- the class attribute, number of contracts, was treated as discrete.

This process resulted in 387 numeric and nominal, and 797 string attributes.

Yonsei University (2014 participant, [10]) made the classification based on nine features present both in the training and testing dataset. The number of attributes is much smaller than that of [18], since only the properties from the Public Contracts Ontology were used (and not those from the vocabularies of interlinked resources, such as the DBpedia ontology), and only those having a low number of missing values.

University of Amsterdam (2013 participant, [6]) used a relational miner based on graph kernels. Pre-processing to tabular format was thus not needed. For the SVM classifier, the class attribute (number of contracts) was transformed to intervals according to the statistical distribution of the values of the *numberOfTenders* property.

5.3 Pre-Processing of Polish Data

The main task was enrichment of the collection. First, addresses in public procurement data do not contain information about *powiat* (district, administrative unit smaller than voivodeship). In order to add this attribute we had to derive data from another database. By using Silk (LOD2 stack tool) we integrated our dataset with part of TERYT Vocabulary (Polish geographical datasource containing list of territorial division units, localities, streets and dictionary of symbols and codes).

Accurate locations related to organisations can be useful for geographical analysis of public procurement. We utilised Nominatim[51] as a tool to extract geocoding points – it returns several coordinates in response to an address query. We had to elaborate heuristic to process outcome because used external tool sometimes returned multiple inconsistent results. As a result we assigned geocoding points about organisations to procurement collection.

For improving the interoperability of the contracts between countries, several codes were introduced. CPV is one of them. Codes from procurement linked to descriptions appends meaning to their raw representation, allowing better categorisation and interpreting them by machines. The detailed CPV vocabulary is available in on-line Official Journal of the European Union (*commission regulation number 213/2008*) as PDF file. Codes

[50] This means, in the given RDF serialization, i.e. more-or-less random selection.
[51] http://wiki.openstreetmap.org/wiki/Nominatim

with descriptions extracted to CSV file, then transformed to triples in similar way as public contracts (mentioned above) and linked to appropriate inscription from contracts with their description.

6 Choice of Tools

At this point we do not specify whether the techniques should be applied to original linked data or to the data converted into table format – both approaches are possible [5]. Some interesting approaches in the latter group include FeGeLOD [9], more recently reimplemented as a RapidMiner tool extension[52] which offers the functionality of machine learning from linked data for a given dataset in an automatic way. Similarly, RMonto – Semantic Data Mining[53] – uses RapidMiner as its basis. On the other end we have methods dealing purely with linked data: intersection of tree path kernel or Weisfeiler-Lehman graph kernel [6]. Direct access to triples is done via SPARQL – there is an extension SPARQL-ML offering statistical relational learning methods, e.g., relational similarity trees and Bayes classifiers.

Specific problems that are intrinsic for graph data include: unbalanced learning data (different counts of classes), overlap of the classes, multidimensionality, big number of potential attributes, and loss of information during transformation from graph to tabular data.

We start the description of tools with visual tools, for which we already show examples of use. For data mining tools proper, we defer the presentation of results to Section 7.

6.1 I2G Visualization of Statistics on Datasets

For Polish data, two web applications for data visualization have been developed. It is important to note that they rely purely on RDF data (accessing it via SPARQL) and therefore lay the foundations for improved usability of linked data.

The first application shows statistics of public contracts in Poland on charts. Application is divided into three main sections: general (simple statistics e.g., number of all contracts, lack of notice award), geographic (e.g., number of contracts by province, number of contracts by city, average estimated value of contract by province) and business entities (e.g., number of suppliers, number of rejected offers by entity, see fig. 1, number of contract notice by type of entity).

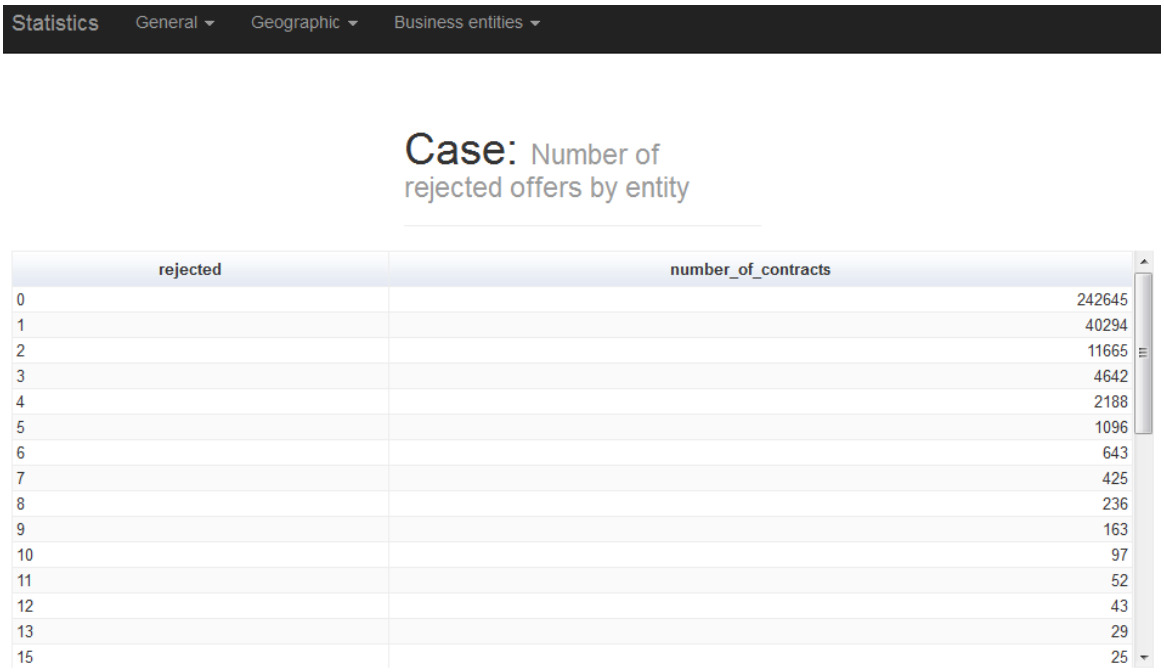
Figure 2 shows number of suppliers of top contractors. Figure 3 presents pie charts characterising popularity of contract type and procedure type.

Another application presents Polish aggregate statistics by geographical dimensions reflecting administrative division of Poland. It was developed based on the procurement collection enrichment by Polish geographical datasets. Contour maps were extracted from Open Street Map.[54] All measures are available in the form of drill-down maps via graphical interface. There are two levels of map: province (see fig. 4) and district (see fig. 5).

6.2 Visual Exploration of RDF Graphs – Payola

Payola is a web application that allows to create RDF data processing pipelines and connect them with visualizer plugins from a large library. One of its modes of operation is the use of SPARQL CONSTRUCT queries over endpoints to create a link/node graph. Entities in the graph can then be automatically arranged via specific

[52] http://dws.informatik.uni-mannheim.de/en/research/rapidminer-lod-extension
[53] http://www.e-lico.eu/rmonto.html
[54] http://wiki.openstreetmap.org/wiki/



rejected	number_of_contracts
0	242645
1	40294
2	11665
3	4642
4	2188
5	1096
6	643
7	425
8	236
9	163
10	97
11	52
12	43
13	29
15	25

Figure 1: Sample statistics about Polish public contracts presented in a table

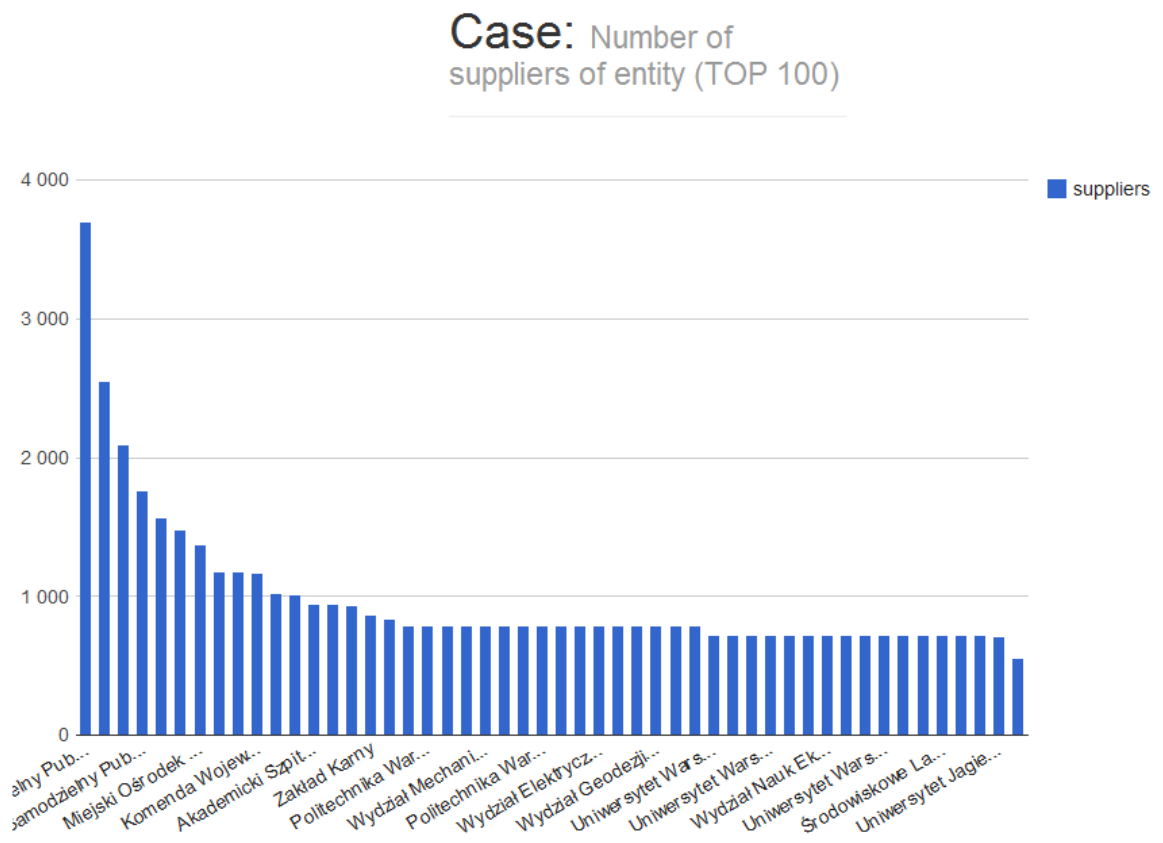


Figure 2: Sample bar chart showing statistics about Polish public contracts

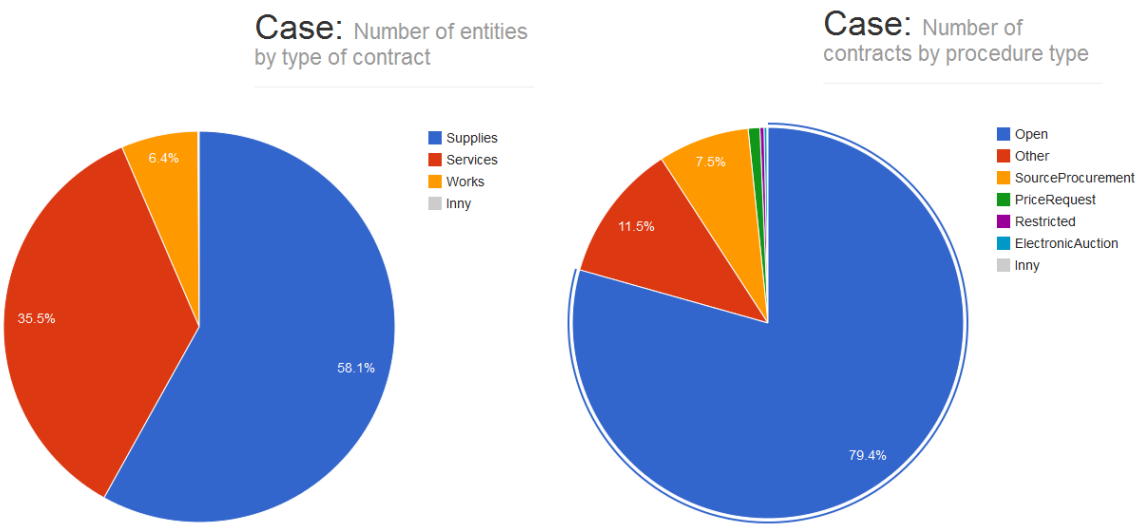


Figure 3: Sample pie charts showing statistics about Polish public contracts

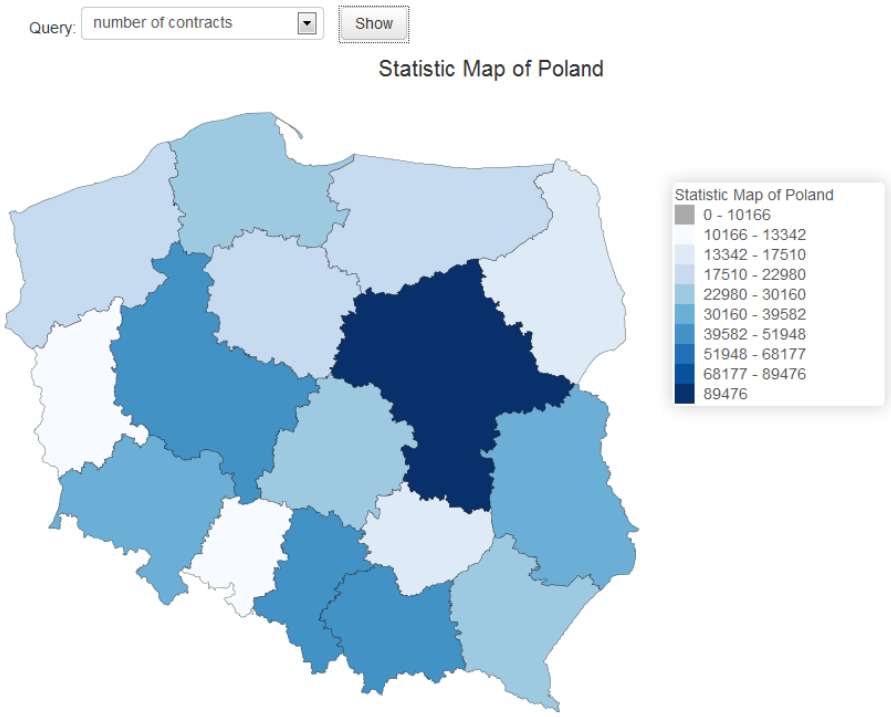


Figure 4: Number of public contracts in Poland in 2013 by province

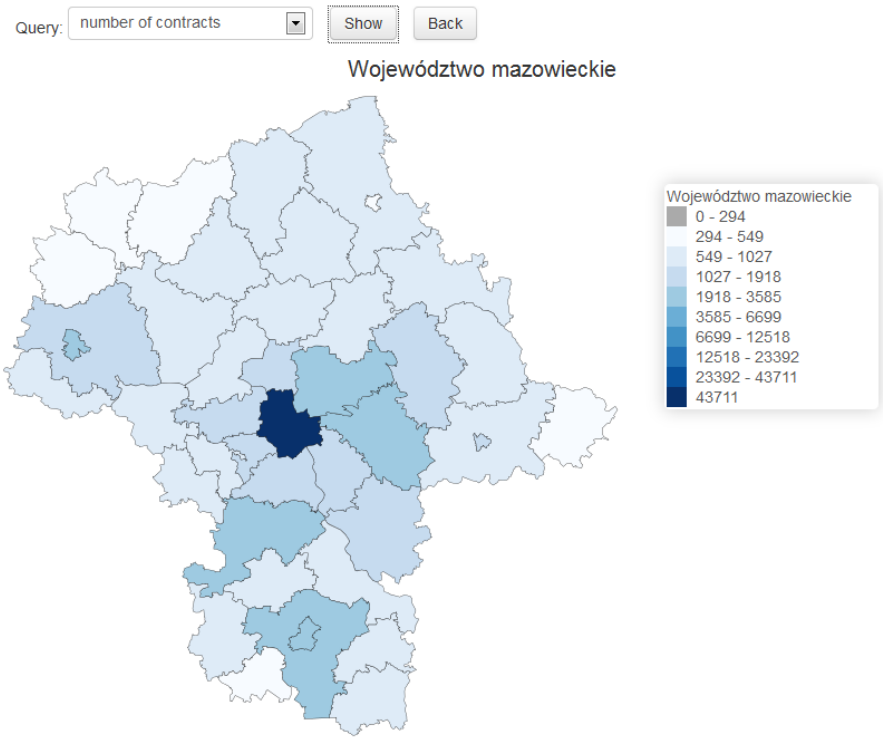


Figure 5: Number of public contracts in Mazovia in 2013 by districts

visualizers (tree, circular, gravity etc.), manually dragged, and customized using coloring and glyphs with respect to ontological types.

We experimented with the link/node graph as possible source of fine-grained insights into the data structures. The graph can, for example, present the bidders together with their year of registration, as in Fig. 6. It may reveal that in the Czech republic multiple companies highly active in procurement bidding were founded in the same year, and allow to immediately inspect the name of the company and its number of contracts. Since the Payola back-end as well as interface is currently undergoing a major rewrite in connection with the introduction of a new DataCube visualization plugin, we however postponed more substantial analyses (comparing/coupling the visualization scenarios with the data mining scenarios) to a later phase of the project.

6.3 Graph Summarization – B-Annot

Graph summarisation displaying the use of various vocabularies is important for broad datasets with unknown structure. This is not entirely the case for our datasets, as they have been transformed to RDF in a transparent way and their schema is directly derived from that of the PCO. However, as it is an analytical task as well, we include, in Fig. 7, the result of frequent path computation in the Czech Bulletin of Public Contracts with respect to PCO, as carried out in the dataset analysis component of B-Annot (vocabulary analysis tool described in detail in Deliverable D4.4a[55] and [27]).

[55] http://lod2.eu/Deliverable/D4.4a.html

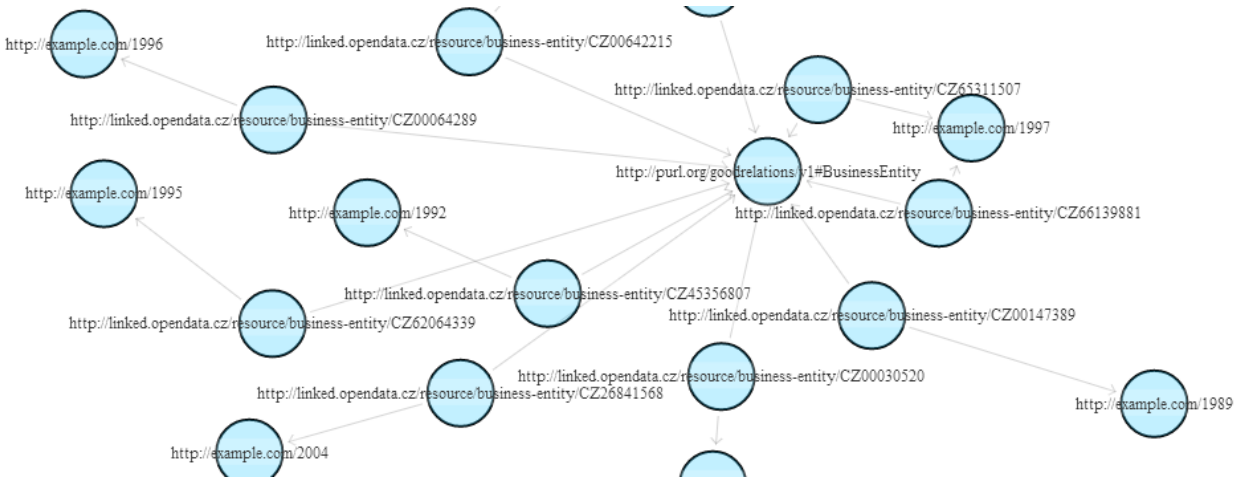


Figure 6: Bidders with years of establishment, in Payola

Summary for dataset

Show paths with occurrences greater than: 0 Show

Type 1	Property	Type 2	Count
adms:Identifier	dcq:creator	gr:BusinessEntity	285562
gr:BusinessEntity	schema:address	schema:PostalAddress	213070
schema:PostalAddress	schema:streetAddress		212800
schema:PostalAddress	schema:addressLocality		212570
schema:PostalAddress	schema:postalCode		211299
schema:PostalAddress	schema:addressCountry		210970
schema:ContactPoint	schema:name		201410
pc:Contract	adms:identifier	adms:Identifier	178928
skos:Concept	skos:inScheme		119430
pc:Contract	pc:contractingAuthority	gr:BusinessEntity	109879
schema:ContactPoint	schema:faxNumber		106326
pc:Contract	pc:awardedTender	pc:Tender	102425
gr:BusinessEntity	schema:contact	schema:ContactPoint	101839
schema:Place	schema:name		96787
schema:ContactPoint	schema:telephone		79438
schema:ContactPoint	schema:email		77321
pc:ContractAwardNotice	rdfs:seeAlso		73776
pc:ContractAwardNotice	adms:identifier	adms:Identifier	73764
pc:ContractNotice	rdfs:seeAlso		32870
pc:ContractNotice	adms:identifier	adms:Identifier	32870
gr:BusinessEntity	pc:activityKind	skos:Concept	29000
schema:ContactPoint	schema:description		28201
pc:TendersOpening	schema:location	schema:Place	28096
gr:BusinessEntity	pc:activityKind		24271
pc:Contract	pc:additionalObject		6048

Figure 7: Frequent Class-Property-Class paths in a Czech procurement dataset

6.4 Mainstream Propositional Data Miners

The strength of the linked data analysis in single-table (in data mining terms, propositional) format is the availability of data mining software with numerous implemented algorithms. Among the most popular tools are nowadays Weka, RapidMiner and SAS Enterprise Miner.

Weka (Waikato Environment for Knowledge Analysis) is collection of machine learning algorithms for data mining tasks, written in Java programming language. It was developed by The University of Waikato, New Zealand on GNU General Public License. There are two different ways of usage: 1. Desktop Application (Data Mining Software in Java), 2. The Weka Java API (Java library). This tool provides several mechanisms for data preprocessing, mining and visualization. It provides implementations of relevant methods of classification (e.g., Naive Bayes, J48, SVM, JRip), clustering (e.g., kNN, EM, Cobweb), associations (Apriori, FPGrowth) relative to study tasks such as looking similar contracts, discover ties between contracting authorities, bidders and offers, prediction number of bidders, prediction successful tender and prediction of multi-contracts. Weka support several file formats including most popular CSV and ARFF.

RapidMiner is software that provides an integrated environment for data mining, machine learning, text mining, predictive analytics and business analytics on AGPL license, written in Java programming language. It supply more implemented methods, intuitive graphical interface, readable visualisation, a lot of plugins and supported file formats compared to Weka. RapidMiner can uses Weka's method (by Weka Extension) but also has a lot of different algorithms interesting in our case, such as DBScan or ID3.

SAS Enterprise Miner is a commercial tool offering standard data mining functionality in a form highly accessible to a business user.

In the initial phase of the research, WEKA was used by I2G, and also by a third party (researchers from University of Darmstadt, as part of their participation in the Linked Data Mining Challenge in 2013). SAS Enterprise Miner was used by a third party (researchers from Yonsei University, Seoul, as part of their participation in the Linked Data Mining Challenge in 2014). The application of RapidMiner was slightly deferred, however, as this system is most RDF-friendly among the propositional miners, it will be systematically applied in the subsequent phase.

6.5 4ft-Miner: Discovery of Rich Associations

We employed the *4ft-Miner* procedure as the most popular procedure of the *LISp-Miner* data mining system [24]. *4ft-Miner* mines for association rules of the form $\varphi \approx \psi/\xi$, where φ, ψ and ξ are called *antecedent*, *succedent* and *condition*, respectively. Antecedent and succedent are conjunctions of *literals*. Literals are derived from attributes, i.e. fields of the underlying data matrix; unlike most propositional mining system, they can be (at runtime) equipped with complex *coefficients*, i.e. value ranges. The association rule $\varphi \approx \psi/\xi$ means that on the subset of data defined by ξ, φ and ψ are associated in the way defined by the symbol $\approx$. The symbol $\approx$, called *4ft-quantifier*, corresponds to some statistical or heuristic test over the four-fold contingency table of φ and ψ.

The task definition language of *4ft-Miner* is thus substantially richer than that of conventional association mining tools. While this complexity may lead to dramatic computational overhead if the tool is used in a naive way, trained analysts can shape powerful queries that stand in between elementary querying (such as using SQL or SPARQL) and large-scale, simply defined data mining (e.g., using the Apriori algorithm).

6.6 Relational and RDF Native Miners

The prime RDF native tool applied on the data was *DL-Learner*, developed by ULEI (now part of the LOD Stack). DL-Learner is a tool for learning concepts in Description Logics (DLs) from user-provided examples. Equivalently, it can be used to learn classes in OWL ontologies from selected objects. The goal of DL-Learner is to support knowledge engineers in constructing knowledge and learning about the data they created. The data mining tasks considered are

- Positive and Negative Examples: finding an OWL class expression C such that all/many positive examples are instances of C w.r.t. a given background ontology O, and no/few negative examples are instances of C w.r.t. O.
- Positive Examples: finding a class expression which closely fits the positive examples while still generalising sufficiently well.
- Class Learning: similar to the previous one, but focusing on short and readable class expressions rather than on covering the examples as precisely as possible.

The second tool applied on the data, this time by a third party (researchers from University of Amsterdam, as part of their participation in the Linked Data Mining Challenge in 2013) was an implementation of *graph kernels*. The approach, described in [6], is based on the idea that RDF instances are represented by their subgraphs. By computing a kernel function on the subgraphs and then training a classifier, i.e. a support vector machine, properties for the instances can be predicted. In the study, SVC classification and SVR regression was used.

7 Data Modeling and Analytics

7.1 Initial Exploration of Polish Data

There are two important problems for data processing that we tackled: missing values and non-normalised text. One could say that they are intrinsic for linked data mining. However, we start with plain data and then lift it to the linked data, so classical data mining issues has to be at least partly addressed. The first problem, missing values, is not so severe, as some values can be imputed from tree structure.

The latter problem is particularly important for usability of the system. Dirty data means that various entities cannot be easily linked, for example matching of company names is not trivial. Such names very often differ just in white spaces, dots, acronyms, sometimes they have attached legal form. Moreover, data is introduced using various conventions. We should also consider removal of some attributes that are not sufficiently discriminant. For this purpose measures like information gain, gain ratio or Chi-square tests can be used. In the case of grouping, we can identify outlier contracts as a by-product, i.e. specific, rare or even suspicious contracts.

Services and supplies are the most popular contracts when considering the type of contract, contributing approximately 64% of all contracts. The most frequent procedure was the open tendering (79.4%). Intriguingly high was the percentage of negotiated contracts, which is the least formal mode of public procurement (7.5 %).

About half of the calls (51.7%) were from 5 out of 16 wealthiest provinces (in terms of GDP per capita 1 - data from 2011), that is Mazovia, Silesia, Lesser Poland, Lower Silesia and Greater Poland. The biggest number of contracts has been announced in the biggest Polish cities, where the unquestioned leader is the capital city Warsaw. It has awarded twice the number of contracts of the second largest city of Kraków. The other distinctive cities are Wrocław and Poznań.

Our attention should be focused on contracts with high number of rejected offers. This can be caused by a strict criteria that are difficult to meet, by criteria favouring certain bidder or for other reasons. Interesting seem to be three auctions, in which the number of rejected tenders reached the value of 298, 245, 111 (in 2013). We may also be interested in tenders with just one bidder, suggesting the possible collusion.

Most of the bidders took part in the procurement type of delivery. Least popular was construction contracts. In areas of low competitiveness there are better chances for certain abuses.

The most active contracting authorities in Poland are government administration, independent public healthcare organisation, public universities and public bodies (totalling in 88% of contracts). Frequent and repetitive orders placed by the same organizations may tend to favour certain bidders.

Representation of business entities in contract award notices is problematic. There are often entities bearing 2-3 different names (sometimes even 5-6). This is due to the ambiguous spelling – names often vary by white space or a period. A record contract was a unit consisting of a total of about 285 various entities performing its contract. On the other hand, about 6196 units were served by no more than 10 providers, 5281 had no more than 5 suppliers and 2348 had only one supplier. The better picture can be provided be detailed analysis in comparison to similar entities (hence importance of clustering also on the contracting authorities set).

7.2 Frequent Associations in Czech and U.S. Data

Several analytical tasks have been defined on Czech and U.S. procurement data resulting from SPARQL SELECT queries, as described in Section 5.1, and they have been submitted to the LISp-Miner system. An example of task setting is in Fig. 10. It addresses a broad range of association hypotheses relating an interval of agreed price (in the RHS, called Succedent) to the time period of the contract, kind of the (subject of the) contract and the type of contract procedure employed (all in the LHS, called Antecedent). The hypotheses are to be of the Above-Average Difference (AAD) kind, i.e., expressing that a certain succedent value is significantly more often present in connection with the antecedent values than it is in connection with other values.

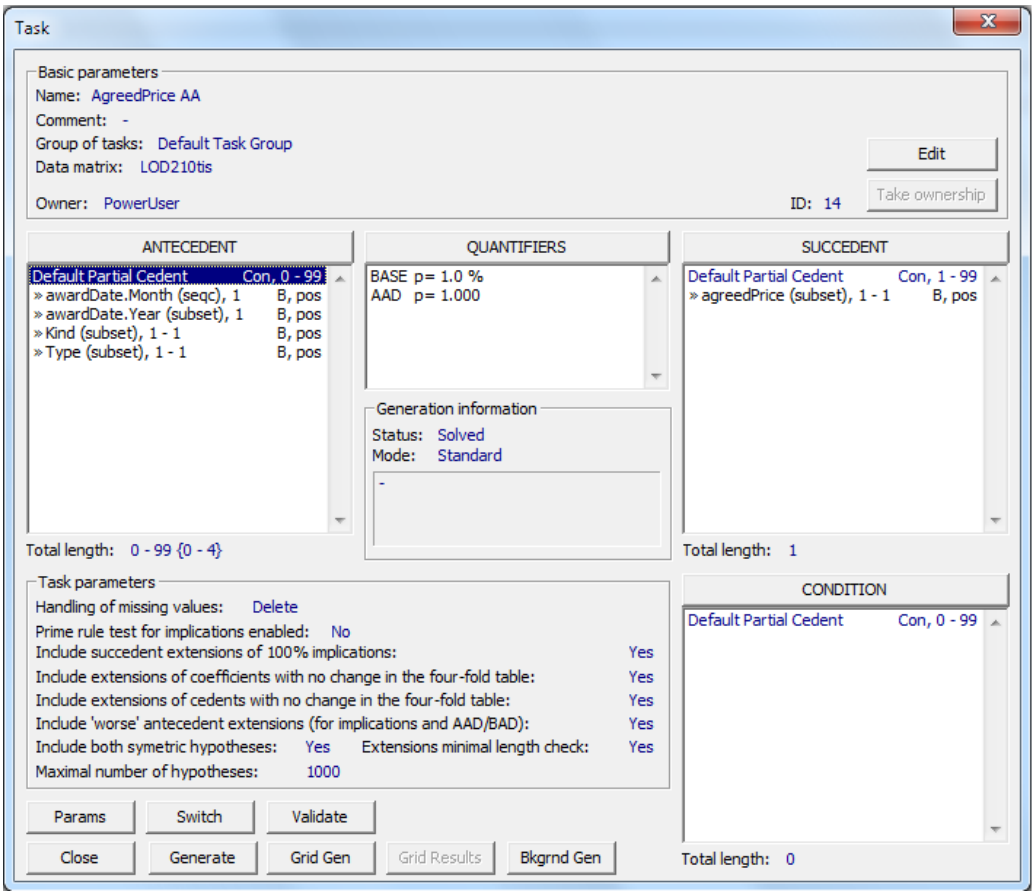


Figure 8: Task setting for LISp-Miner

The results of the data mining run are in Fig. 9. The first, highlighted, association states that very high agreed cost (30M CZK and more) occurs more often when the contract regards works design and the type of the procedure is open. (The value of the quantifier, 1.084, means that the relative frequency of occurrence of this price interval is by 108% per cent, i.e. more than twice, higher for works design with open procedure than is its relative frequency in the set of all contracts.) Note that the set of strong associations does not include the sub-associations having only works design *or* only open procedure in the antecedent. This indicates that the impact on the agreed price may indeed be due to synergic effect of both factors in the antecedent.

Further analyses applied concern, e.g., factors influencing the offered (rather than agreed) price. For example, the summer months make the price of *negotiated* contracts of *works design* type more likely to be in the interval between 10-20M CZK.

1	14	1.084	**Kind**(*WorksDesign*) & **Type**(*Open*) >÷< **agreedPrice**(*<30M;50M)*,*50M+*)
2	13	1.006	**Kind**(*WorksDesign*) & **Type**(*Open*) >÷< **agreedPrice**(*<20M;50M)*)
3	15	0.942	**Kind**(*WorksDesign*) & **Type**(*Open*) >÷< **agreedPrice**(*>=50M+*)
4	17	0.934	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<10000000;20000000>*) / **awardDate.Year**(*2008*)
5	7	0.932	**Kind**(*Supplies*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<300000;5000000>*)
6	11	0.905	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<15000000;20000000>*)
7	12	0.882	**Kind**(*WorksDesign*) & **Type**(*Open*) >÷< **agreedPrice**(*<20M;30M)*)
8	6	0.833	**Kind**(*Supplies*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<1000000;5000000>*)
9	22	0.795	**Kind**(*WorksDesign*) & **Type**(*Open*) >÷< **agreedPrice**(*<5000000;15000000>*) / **awardDate.Year**(*2013*)
10	19	0.780	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<10000000;20000000>*) / **awardDate.Year**(*2009*)
11	10	0.773	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<10000000;20000000>*)
12	5	0.756	**Kind**(*Supplies*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<1000000;3000000>*)
13	23	0.707	**Type**(*Negotiated*) >÷< **agreedPrice**(*<=500000*) / **awardDate.Year**(*2013*)
14	2	0.699	**Kind**(*Supplies*) >÷< **agreedPrice**(*<1000000;3000000>*)
15	9	0.670	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<10000000;15000000>*)
16	4	0.648	**Kind**(*Supplies*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<500000;3000000>*)
17	1	0.635	**Kind**(*Supplies*) >÷< **agreedPrice**(*<500000;3000000>*)
18	24	0.632	**Type**(*Negotiated*) >÷< **agreedPrice**(*<5000;1000000>*) / **awardDate.Year**(*2013*)
19	3	0.589	**Kind**(*Supplies*) >÷< **agreedPrice**(*<1000000;5000000>*)
20	21	0.567	**Kind**(*WorksDesign*) & **Type**(*Open*) >÷< **agreedPrice**(*<300000;10000000>*) / **awardDate.Year**(*2013*)
21	16	0.566	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<5000000;15000000>*) / **awardDate.Year**(*2008*)
22	20	0.548	**Kind**(*Supplies*) >÷< **agreedPrice**(*<500000;3000000>*) / **awardDate.Year**(*2012*)
23	8	0.527	**Kind**(*Supplies*) & **Type**(*Open*) >÷< **agreedPrice**(*<1000000;3000000>*)
24	18	0.516	**Kind**(*WorksDesign*) & **Type**(*Negotiated*) >÷< **agreedPrice**(*<5000000;15000000>*) / **awardDate.Year**(*2009*)

Figure 9: Example results for LISp-Miner: factors influencing the agreed cost in the Czech Republic

For the U.S. data, the analysis focused, among other, on the task of predicting the number of bidders, as formulated in the LDMC task (see also Section 5.2. As an example demonstrating the use of externally fetched LOD cloud data in the antecedent of rules, see a fragment of analysis of U.S. procurement data with respect to the impact various attributes of a contract notice may have on the subsequent number of tenders (Fig. 10). The association rules listed in the table fragment regard both a CPV code of the contract object

1	38	8.753	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*>=1700*) >÷< **Tenders**(*>=50*)
2	39	7.562	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*>=1700*) >÷< **Tenders**(*>=30*)
3	36	7.306	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*>=1600*) >÷< **Tenders**(*>=50*)
4	32	7.190	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1100;1600)*) >÷< **Tenders**(*>=50*)
5	29	7.049	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1000;1500)*) >÷< **Tenders**(*>=50*)
6	37	6.490	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*>=1600*) >÷< **Tenders**(*>=30*)
7	24	6.481	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<800;1300)*) >÷< **Tenders**(*>=50*)
8	33	6.064	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1100;1600)*) >÷< **Tenders**(*>=30*)
9	1	6.022	**mainObject**(*Research and Development in the Physical, Engineer*) >÷< **Tenders**(*N/A*)
10	26	5.944	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<900;1400)*) >÷< **Tenders**(*>=50*)
11	2	5.882	**mainObject**(*Research and Development in the Physical, Engineer*) >÷< **Tenders**(*>=100+*)
12	31	5.866	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1100;1500)*) >÷< **Tenders**(*>=30*)
13	30	5.741	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1000;1500)*) >÷< **Tenders**(*>=30*)
14	28	5.695	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1000;1400)*) >÷< **Tenders**(*>=30*)
15	3	5.571	**mainObject**(*Research and Development in the Physical, Engineer*) >÷< **Tenders**(*>=50*)
16	18	5.494	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<400;900)*) >÷< **Tenders**(*>=50*)
17	16	5.388	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<400;700)*) >÷< **Tenders**(*>=30*)
18	17	5.235	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<400;800)*) >÷< **Tenders**(*>=30*)
19	19	5.210	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<400;900)*) >÷< **Tenders**(*>=30*)
20	20	5.080	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<500;900)*) >÷< **Tenders**(*>=30*)
21	25	5.049	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<800;1300)*) >÷< **Tenders**(*>=30*)
22	34	5.047	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1200;1700)*) >÷< **Tenders**(*>=30*)
23	35	4.965	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<1300;2000)*) >÷< **Tenders**(*>=30*)
24	14	4.923	**mainObject**(*Research and Development in the Physical, Engineer*) & **populationDensityPerKm2**(*<300;800)*) >÷< **Tenders**(*>=50*)
25	4	4.916	**mainObject**(*Research and Development in the Physical, Engineer*) >÷< **Tenders**(*>=30*)

Figure 10: Example results for LISp-Miner: factors correlated with number of tenders in the U.S.

(`mainObject` attribute), originating from one of the core procurement dataset, and the population density attribute, originating from DBpedia. It indicates that contracts for 'Research and Development in the Physical, Engineering, and Life Sciences' in localities with higher population density tend to attract a high number of tenders.

7.3 Predictive Mining Results

The distinguished predictive mining tasks have, to date, only been addressed by third parties, i.e. the LDMC participants. The results are in the respected papers. [18] applied a wide range of WEKA classifiers – J48 decision trees, JRip rules, k-NN, LibSNV and LibLinear – some of them requiring binarization of features. Best results were achieved using an ensemble of classifiers. [10] only applied the decision tree classifier from SAS. Finally, [6] applied their own relational classifier based on graph kernels. They were the only to address not only the prediction of the number of contracts but also the prediction of multi-contracts.[56]

Predictive mining by LOD2 partner tools is ongoing and will be part of D9a.3.2, both in terms of implemented prototype and reported results.

7.4 DL-Learner Results

DL-Learner was applied on the task of distinguishing between successful and unsuccessful tenders. The fact that information on unsuccessful tenders is rarely available led to the choice of the profile feeds of Czech contracting authorities as the source dataset. In the learning task, successful tenders represented positive examples and unsuccessful tenders represented negative examples, in the 'pos-neg' task.

The DL-Learner's (most expressive) CELOE algorithm [17] was applied on a sample of 100 positive and 200 negative examples. The runtime was fixed to 300s. During the mining, the maximum noise allowed had to be gradually decreased downto 50% in order to find some hypotheses. The list of top 10 hypotheses is in Listing 1. It is obvious that the 'descriptions of successful contracts' generated only contain the standard features present for all contracts (positive as well as negative) that have all the required information complete.

```
1: pc:offeredPrice some gr:PriceSpecification (pred. acc.: 51.00%, F-measure:
   67.11%)
2: gr:hasPriceSpecification some Thing (pred. acc.: 51.00%, F-measure: 67.11%)
3: gr:hasPriceSpecification some gr:PriceSpecification (pred. acc.: 51.00%, F-
   measure: 67.11%)
4: (foaf:Person or pc:offeredPrice some gr:PriceSpecification) (pred. acc.:
   51.00%, F-measure: 67.11%)
5: (foaf:Person or gr:hasPriceSpecification some Thing) (pred. acc.: 51.00%, F-
   measure: 67.11%)
6: (foaf:Person or gr:hasPriceSpecification some gr:PriceSpecification) (pred.
   acc.: 51.00%, F-measure: 67.11%)
7: (http://www.w3.org/ns/regorg#RegisteredOrganization or pc:offeredPrice some gr
   :PriceSpecification) (pred. acc.: 51.00%, F-measure: 67.11%)
8: (http://www.w3.org/ns/regorg#RegisteredOrganization or gr:
   hasPriceSpecification some Thing) (pred. acc.: 51.00%, F-measure: 67.11%)
9: (http://www.w3.org/ns/regorg#RegisteredOrganization or gr:
   hasPriceSpecification some gr:PriceSpecification) (pred. acc.: 51.00%, F-
   measure: 67.11%)
10: (http://www.w3.org/ns/adms#Identifier or pc:offeredPrice some gr:
   PriceSpecification) (pred. acc.: 51.00%, F-measure: 67.11%)
```

Listing 1: Top results for DL-Learner

Besides, the task with *positive examples only* was also tried, for completeness. Based on n successful tenders used as positive examples, and a maximum depth of the downloaded tree for each tender is k, the Least General Generalization (LGG) again generated obvious features that all tenders have in common: they

[56] Only presented as task in 2013.

are from a supplier that is located in the Czech Republic, has a legal name and address, and include a price specification that has as currency CZK.

We made a preliminary general conclusion that current DL-based inductive systems are not quite suitable for linked datasets in which the 'richness' is in the distribution of end-values (literals) while the schema is small, simple (even if not completely flat) and homogeneous. It is likely that many current linked data resources (with compact RDB origin) fall under this category. This distinguishes them from linked data based on crowdsourced structures with rich structure of heterogeneous links (properties) between instances of different classes. An example is DBpedia, on which DL approaches previously proved successful.

8 Conclusions and Future Plans

The deliverable summarizes a number of partial studies carried out by UEP, I2G and ULEI, as well as by third parties engaged via two editions of the Linked Data Mining Challenge, between Summer 2013 and Spring 2014. It covers various types of analytical tasks and methods possibly applicable on public procurement data in the form of linked data.

While some of the methods explore RDF data proper, other need it to be converted (back) to a more structured format, essentially, CSV. The complex underlying pipeline appears, for the moment, hard to maintain and error-prone. However, we believe that in long term (assuming the RDF data ETL and linking technology to slightly mature) the benefits of this process will outweigh its costs. Namely, RDF, even if only used as middle-product, has excellent capacity to support interlinking, entity disambiguation, geocoding, as well as flexible construction of various (tabular or graph-shaped) views on data via SPARQL queries. It thus brings clear added value compared to applying data mining merely on the original tabular (or tree-structured, e.g., XML) data in their 'silos'.

Some of these studies will be further extended in the remaining few months of the project, in order to get better insights into the general problem (in many cases, as pioneering research with respect to linked data analytics as such rather than just to procurement linked data analytics). However, the main focus will be the integration of the analytical functionality into the *Public Contracts Filing Application* (PCFA), the early version of which has been described in D9a.1.2.

In the PCFA, the analytical functionality will complement the *matchmaking* functionality that is currently being described in D9a.2.2, in particular, from the point of view of a contracting authority. For a procurement notice under preparation (at least partially filled), it is now possible to find

- similar notices/contracts in the past, from which the remaining information can be copied (or at least getting a hint)
- potentially relevant suppliers, which should learn to know about the existence of the notice.

Further, more *analytical* features then should allow to:

- *Interactively explore*, in graphical form (using the Payola link/node graphs with specific customizations for display of different classes and properties), the linked data about
  - the current notice
  - a (matching) historical notice/contract in graphical form (using Payola)
  - a relevant supplier, including its contracts.
- View *suggested values* for the remaining pieces of contract notice information based on the already provided ones. Association mining (using LISp-Miner in the background) seems to be a suitable technology for that.
- Get an estimate of the *number of bidders* for (as complete as possible) contract notice information. For this, predictive mining as outline in this deliverable is adequate.

The version of PCFA with this functionality implemented is the main target of the forthcoming D9a.3.2.

References

[1] Ziawasch Abedjan and Felix Naumann. Context and target configurations for mining rdf data. In *Proceedings of the 1st International Workshop on Search and Mining Entity-relationship Data*, SMER '11, pages 23–24, New York, NY, USA, 2011. ACM.

[2] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. of the 20th VLDB Conference*, pages 487–499, 1994.

[3] Jose María Alvarez and José Emilio Labra. Semantic methods for reusing linking open data of the european public procurement notices. In *ESWC PhD Symposium*, 2011.

[4] Jose María Alvarez, José Emilio Labra, Ramón Calmeau, Ángel Marín, and Jose Luis Marín. Innovative services to ease the access to the public procurement notices using linking open data and advanced methods based on semantics. In *5th International Conference on Methodologies, Technologies and enabling eGovernment Tools*, 2011.

[5] Claudia d'Amato, Petr Berka, Vojtěch Svátek, and Krzysztof Węcel, editors. *Proceedings of the International Workshop on Data Mining on Linked Data, with Linked Data Mining Challenge collocated with the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECMLPKDD 2013)*, volume 1082 of *CEUR Workshop Proceedings*, Prague, Czech Republic, 2013. CEUR-WS.org.

[6] Gerben Klaas Dirk de Vries and Steven de Rooij. A Fast and Simple Graph Kernel for RDF. In d'Amato et al. [5].

[7] I. Distinto, M. d'Aquin, and Motta E. LOTED2: an Ontology of European Public Procurement Notices. Under review for *Semantic Web – Interoperability, Usability, Applicability*, IOS Press, 2014.

[8] Vladimir Estivill-Castro. Why so many clustering algorithms: A position paper. *SIGKDD Explor. Newsl.*, 4(1):65–75, June 2002.

[9] Daniel Hienert, Daniel Wegener, and Heiko Paulheim. Automatic Classification and Relationship Extraction for Multi-Lingual and Multi-Granular Events from Wikipedia. In *Detection, Representation, and Exploitation of Events in the Semantic Web (DeRiVE 2012)*, volume 1082 of *CEUR Workshop Proceedings*, pages 1–10. CEUR-WS.org, 2012.

[10] Dongkyu Jeon and Wooju Kim. Development of prediction model for linked data based on the decision tree – for track a, task a1. In *Know@LOD 2014, Linked Data Mining Challenge paper*, 2014.

[11] D. Kavitha, B. V. Manikyala Rao, and V. Kishore Babu. A Survey on Assorted Approaches to Graph Data Mining. *International Journal of Computer Applications*, 14(1):43–46, 2011.

[12] Mansoor Ahmed Khan, Gunnar Aastrand Grimnes, and Andreas Dengel. Two pre-processing operators for improved learning from SemanticWeb data. In *First RapidMiner Community Meeting And Conference (RCOMM 2010)*, volume 20, 2010.

[13] C. Kiefer, A. Bernstein, and M. Stocker. The Fundamentals of iSPARQL: A Virtual Triple Approach for Similarity-Based Semantic Web Tasks. In *The Semantic Web*, 2008.

[14] Christoph Kiefer, Abraham Bernstein, and André Locher. Adding Data Mining Support to SPARQL Via Statistical Relational Learning Methods. In Sean Bechhofer, Manfred Hauswirth, Jörg Hoffmann, and Manolis

Koubarakis, editors, *The Semantic Web: Research and Applications*, volume 5021 of *Lecture Notes in Computer Science*, pages 478–492. Springer Berlin Heidelberg, 2008.

[15] Jakub Klímek, Tomáš Knap, Jindřich Mynarz, Martin Nečaský, and Vojtěch Svátek. LOD2 deliverable 9a.1.1: Framework for creating linked data in the domain of public sector contracts. Technical report, LOD2, Prague, 2012.

[16] Krys J. Kochut and Maciej Janik. SPARQLeR: Extended SPARQL for Semantic Association Discovery. In *Proc. of the 4th European Semantic Web Conference (ESWC*, pages 145–159, 2007.

[17] Jens Lehmann. DL-Learner: learning concepts in description logics. *Journal of Machine Learning Research (JMLR)*, 10:2639–2642, 2009.

[18] Eneldo Loza Mencia, Simon Holthausen, Axel Schulz, and Frederik Janssen. Using data mining on linked open data for analyzing e-procurement information - a machine learning approach to the linked data mining challenge 2013. In Claudia d'Amato, Petr Berka, Vojtěch Svátek, and Krzysztof Węcel, editors, *DMoLD*, volume 1082 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[19] Jan Michelfeit and Tomáš Knap. Linked Data Fusion in ODCleanStore. In Birte Glimm and David Huynh, editors, *International Semantic Web Conference (Posters and Demos)*, volume 914 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2012.

[20] Heiko Paulheim. Exploiting Linked Open Data as Background Knowledge in Data Mining. In d'Amato et al. [5].

[21] Heiko Paulheim and Johannes Fürnkranz. Unsupervised Generation of Data Mining Features from Linked Open Data. In *Proceedings of the 2Nd International Conference on Web Intelligence, Mining and Semantics*, WIMS '12, pages 31:1–31:12, New York, NY, USA, 2012. ACM.

[22] J. Potoniec and A. Ławrynowicz. RMonto: Ontological extension to Rapid-Miner. In *Poster and Demo Session of the ISWC 2011, 10th International Semantic Web Conference*, Bonn, Germany, 2011.

[23] Jan Rauch. *Observational Calculi and Association Rules*, volume 469 of *Studies in Computational Intelligence*. Springer, 2013.

[24] Jan Rauch and Milan Šim□unek. An alternative approach to mining association rules. In Tsau Young Lin, Setsuo Ohsuga, Churn-Jung Liau, Xiaohua Hu, and Shusaku Tsumoto, editors, *Foundations of Data Mining and knowledge Discovery*, volume 6 of *Studies in Computational Intelligence*, pages 211–231. Springer, 2005.

[25] Jose María Alvarez Rodríguez, José Emilio Labra Gayo, Ramón Calmeau, Ángel Marín, and Jose Luis Marin. Query expansion methods and performance evaluation for reusing linking open data of the european public procurement notices. In *Proceedings of CAEPIA 2011*, 2011.

[26] Jose María Álvarez Rodríguez, José Emilio Labra Gayo, Francisco Adolfo Cifuentes Silva, Giner Alor-Hernández, Cuauhtémoc Sánchez, and Jaime Alberto Guzmán Luna. 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 (IJSEKE)*, 22(3):365–384, 2012.

[27] Vojtěch Svátek, Simone Serra, Miroslav Vacura, Martin Homola, and Jan Kluka. B-annot: Supplying background model annotations for ontology coherence testing. In Patric Lambrix, Guilin Qi, Matthew Horridge, and Bijan Parsia, editors, *WoDOOM 2014*, CEUR Workshop Proceedings. CEUR-WS.org, 2014.

[28] Vojtěch Svátek, Jindřich Mynarz, and Petr Berka. Linked Data Mining Challenge (LDMC) 2013 summary. In d'Amato et al. [5].

[29] F. Valle, M. d'Aquin, T. Di Noia, and E. Motta. LOTED: Exploiting Linked Data in Analyzing European Procurement Notices. In *1st Workshop on Knowledge Injection into and Extraction from Linked Data collocated with EKAW'10*, Madrid, Spain, 2010. CEUR-WS.org.