Open Data and Data Analysis

Big Data and Artificial Intelligence Approach to Open Data

Matej Kovačič

matej.kovacic@ijs.si

Jožef Stefan Institute

**Centre for Knowledge Transfer
in Information Technologies**

Artificial Intelligence Laboratory

Open data

Open data are data that are freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control.

Elements of open data:
- Availability and access: available as a whole, free or for reasonable reproduction costs, preferably downloadable from internet, available in a convenient and modifiable form.
- Re-use and redistribution: terms of use permit re-use and redistribution.
- Universal participation: for anyone for any purpose (no copyright, patents, other control mechanisms)

Formally, open data has a licence saying it is open data. Sometimes licence requires attribution and share-alike.

Open data

Tim Berners-Lee, the inventor of the Web and Linked Data initiator, suggested a 5-star deployment scheme for Open Data.

Source: http://5stardata.info/
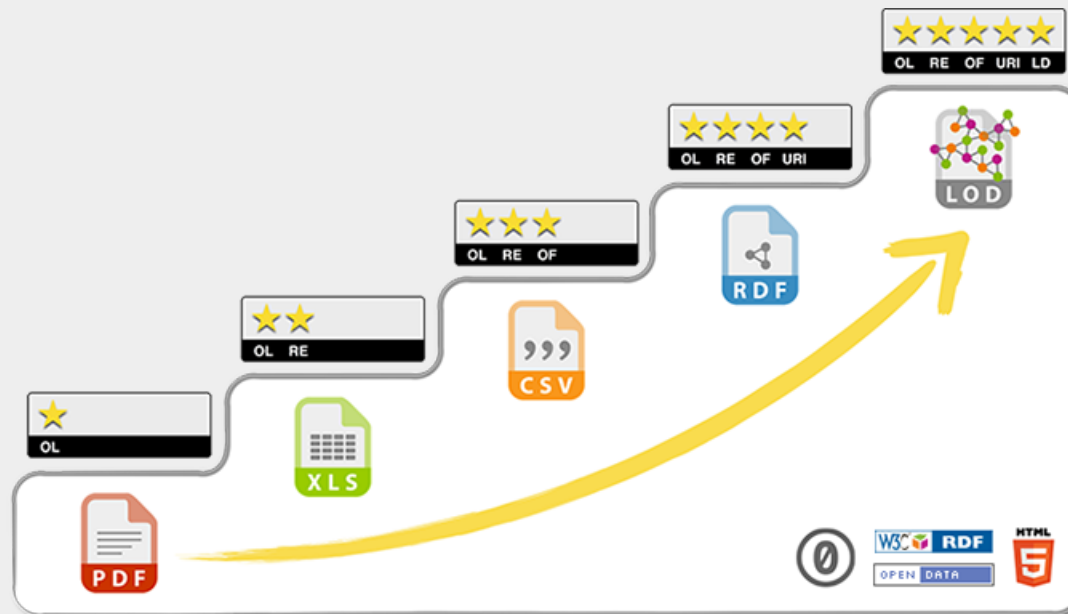
1. Make your data available on the Web (in whatever format) under an open license.
2. Make it available as structured data (e.g., Excel instead of image scan of a table).
3. Make it available in a non-proprietary open format (e.g., CSV instead of Excel.
4. Use URIs to denote things, so that people can point at your data.
5. Link your data to other data to provide context.

Public information

Public information is all information originating from the field of work of the public sector bodies and occurring in the form of a document, a case, a dossier, a register, a record or other documentary material (hereinafter referred to as "the document") drawn up by the body, by the body in cooperation with other body, or acquired from other persons.

Natural and legal persons can under certain conditions use this information also for commercial or non commercial reuse.

Some sources of open data

- Open data in science:
  https://en.wikipedia.org/wiki/Open_data#In_science

- Open data in government:
  https://en.wikipedia.org/wiki/Open_data#In_Government

- EU Open data portals:
  https://ec.europa.eu/digital-single-market/en/open-data-portals

- Open data by American Government:
  https://www.data.gov/

- Slovenian National Interoperability Framework Portal:
  http://nio.gov.si/nio/vstopna.nio?lang=en

- Slovenian community Opendata.si: https://opendata.si/

Big Data

Big data is data sets that are so large or complex that traditional data processing application software is inadequate to deal with them.

Big data challenges include capturing data, data storage, data analysis, search, sharing, transfer, visualization, querying, updating and information privacy.

Lately, the term "big data" tends to refer to the use of predictive analytics, user behavior analytics, or certain other advanced data analytics methods that extract value from data, and seldom to a particular size of data set.

Example of Big Data analysis: http://eventregistry.org/

Big Data

Big data are usually raw, "found" or "organic data". Usually in real time.

Machine (for example from sensors) or user generated.

In social sciences: data about behavior in the natural environment and in real time.

Problems with big data for research purposes:
- Secondary data, intended for another primary use;
- Sometimes not structured or they are not well defined;
- Data quality? Cheap to collect, but expensive to clean and process;
- Ethical concerns: no informed consent policies, privacy issues.

Some examples of open data and its analysis

Slovenian business registry

Ajpes (The Agency of the Republic of Slovenia for Public Legal Records and Related Services) provided:
- ePRS Slovenian Business Register (central public database on all business and legal entities, their subsidiaries, and other organizations` segments located in Slovenia (public and private institutes)), including deleted companies;
- eRTR – Register of legal entity`s bank accounts.

Some other basic registers...

- *Registry of budget users*: list and classification of all direct and indirect Slovenian budget users.

- *Standard Classification of Activities*: the obligatory national standard used for recording, collecting, processing, analysing, mediating and disseminating data connected with activity).

- *CPV (Common procurement vocabulary)*: CPV establishes a single classification system for public procurement aimed at standardising the references used by contracting authorities and entities to describe the subject of procurement contracts.

- ...

Bad investments transferred on *Bank Assets Management Company*

The Bank Assets Management Company (BAMC) was established in March 2013 as a government-owned company with the task of facilitating the restructuring of banks with systemic importance that were facing severe solvency and liquidity problems.

Non-performing assets of commercial banks had been transferred to BAMC.

Bad investments transferred on *Bank Assets Management Company*

Data has been imported into the database and linked to business registry...
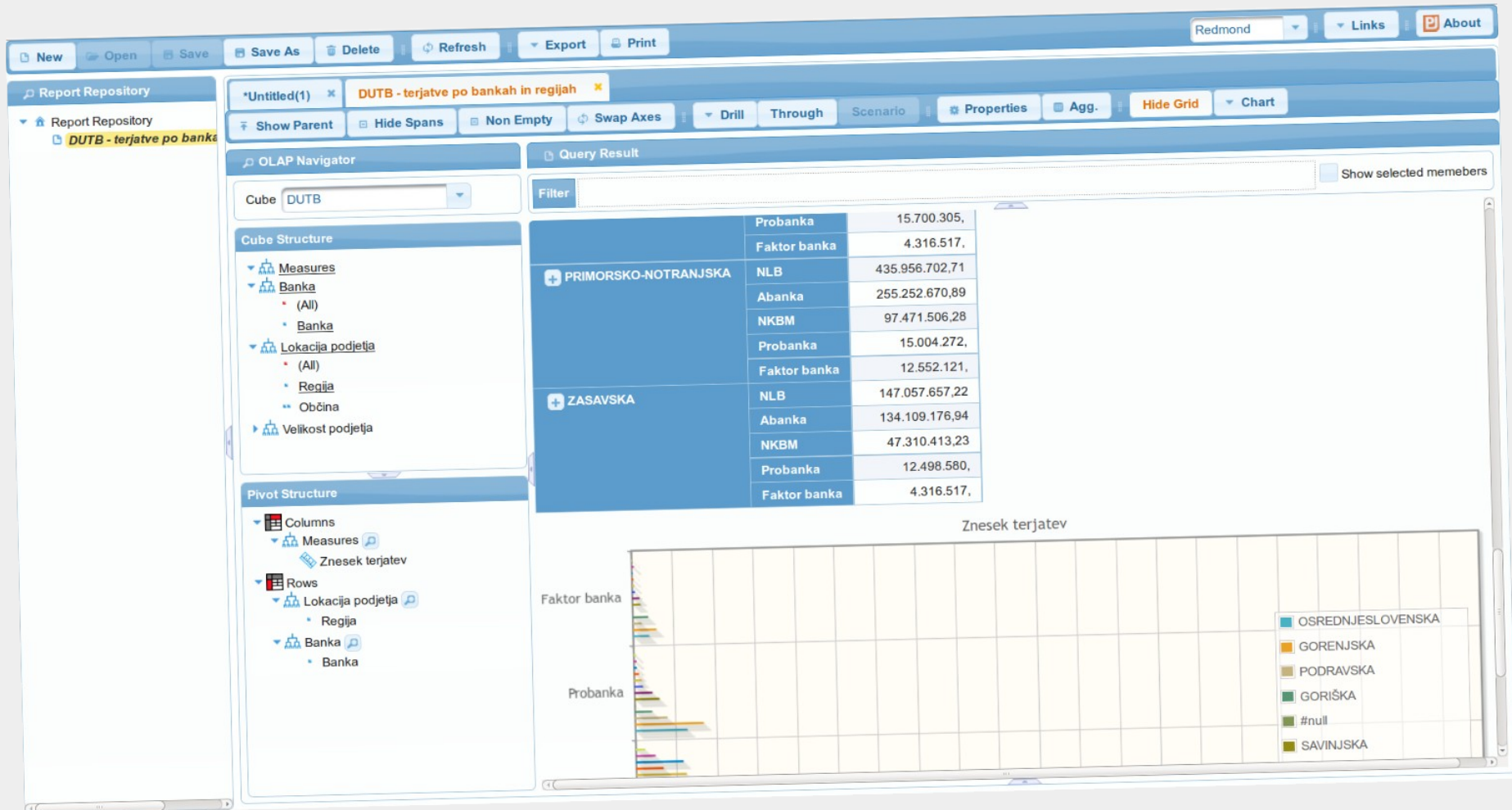
… multiple bad loans to companies transferred to BAMC has been visualised…
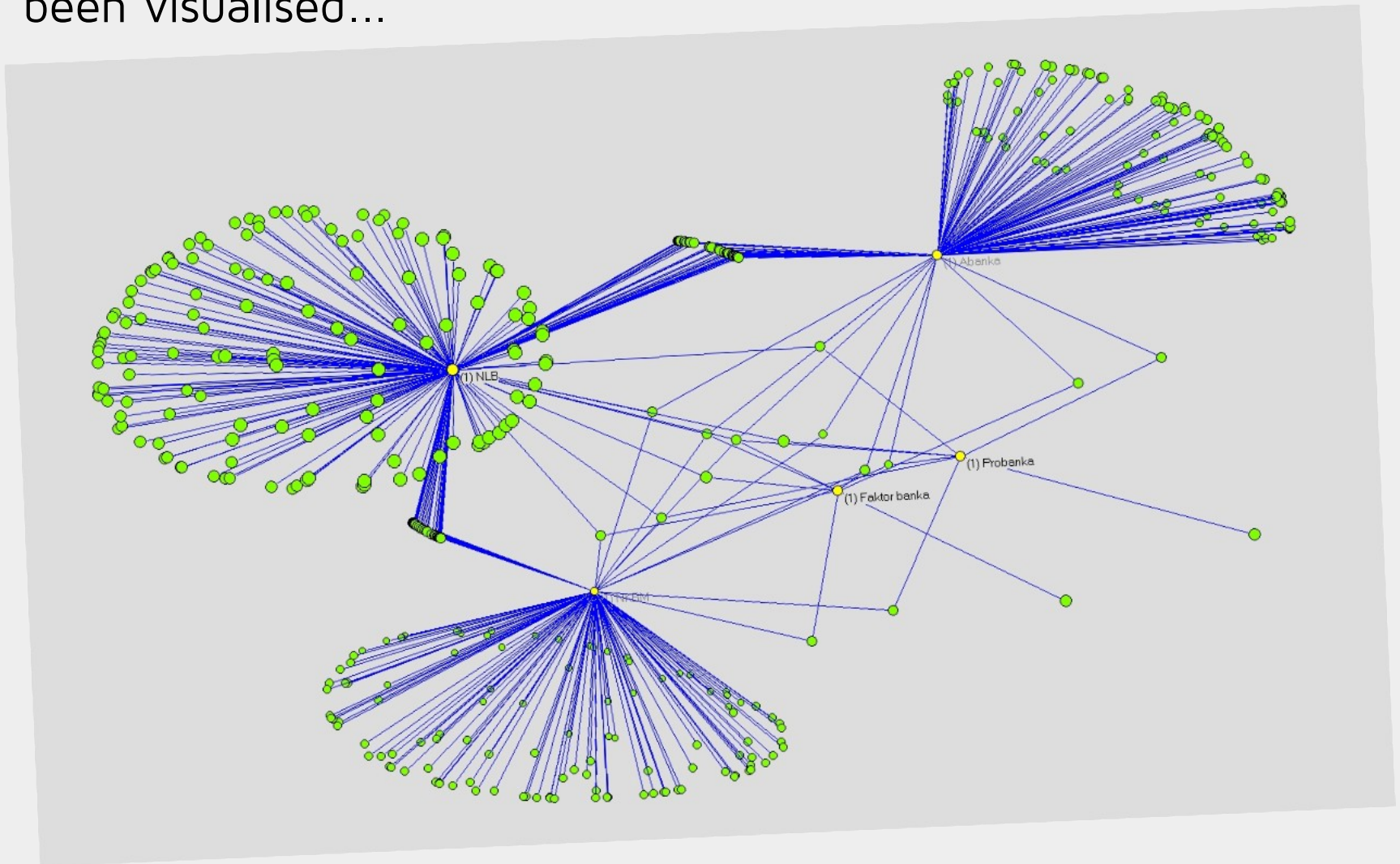
… the network of connected persons has been created (persons who are or have been directors and/or members of supervisory boards of those companies.
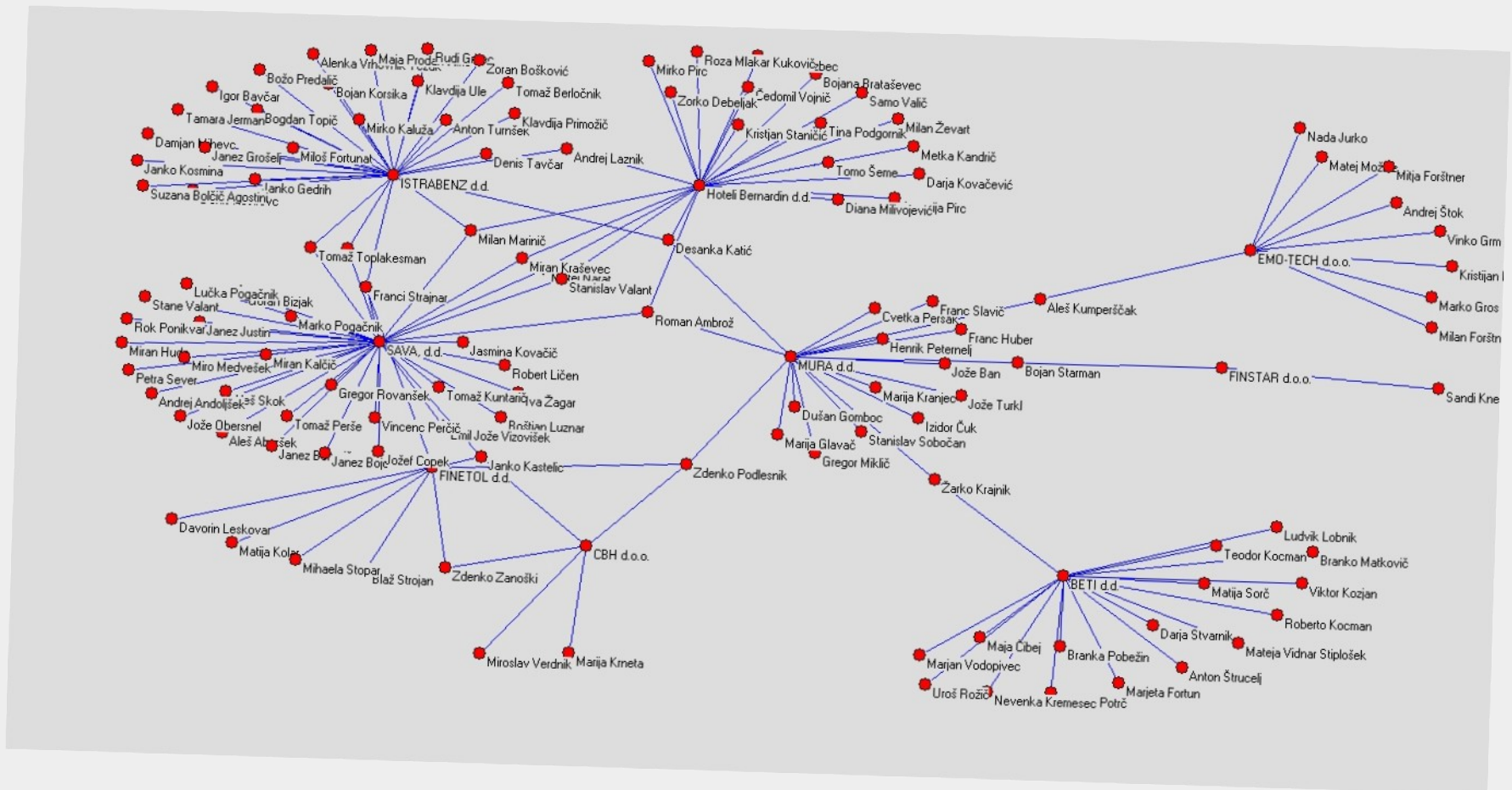
The network shows (management) connections among corporations…
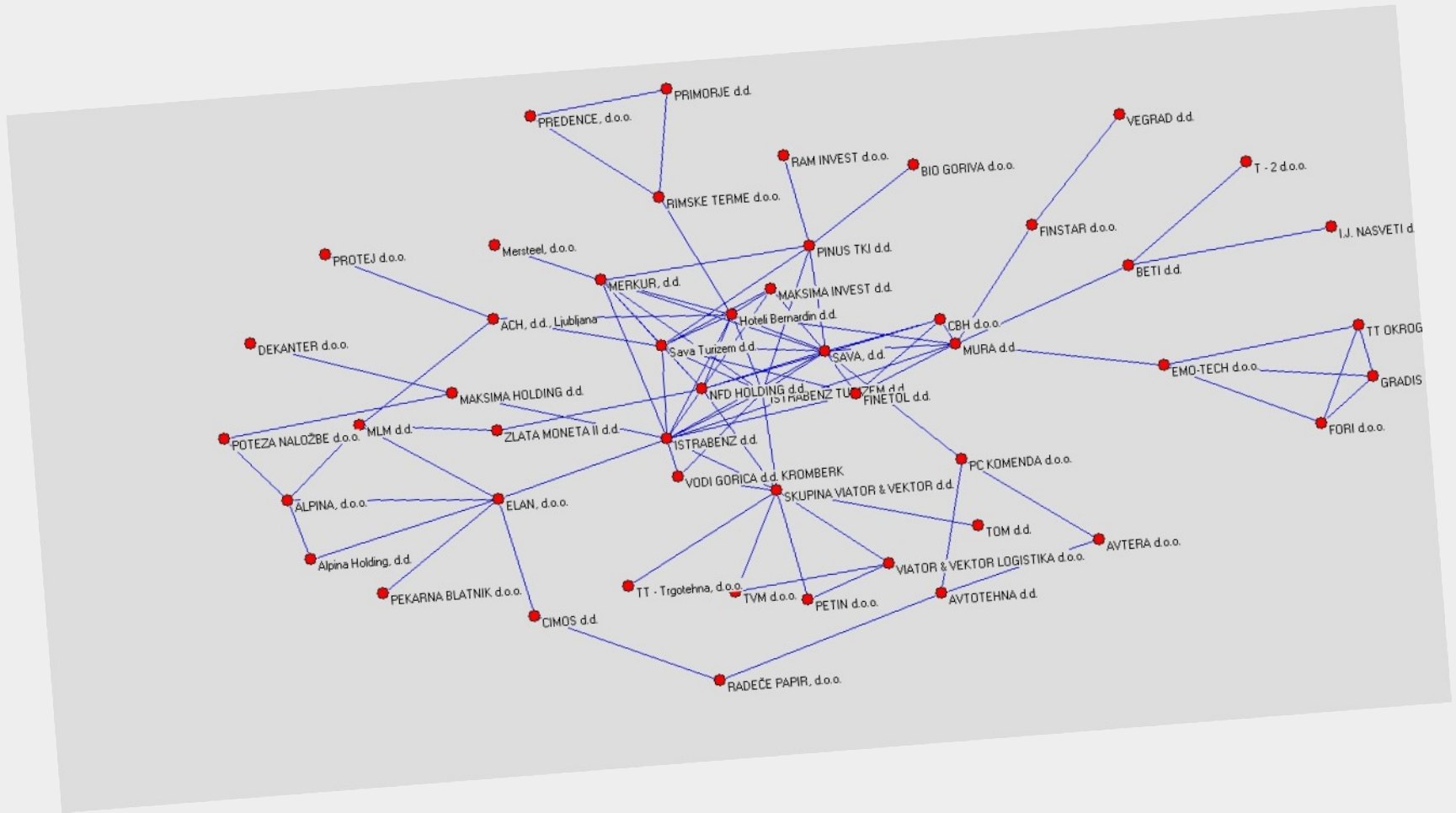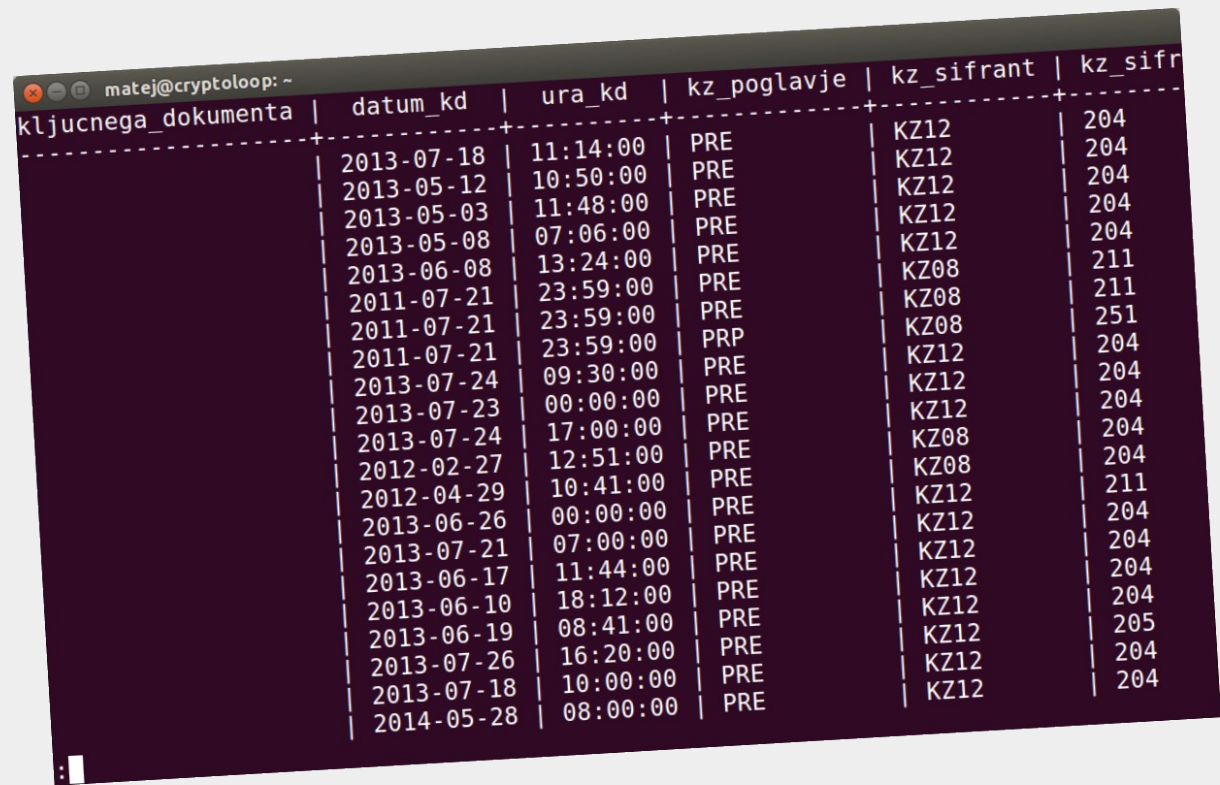
The biggest component of the network contains approx. 33% of all companies transferred to BAMC. Those companies has been granted approx. 61% of bad loans.
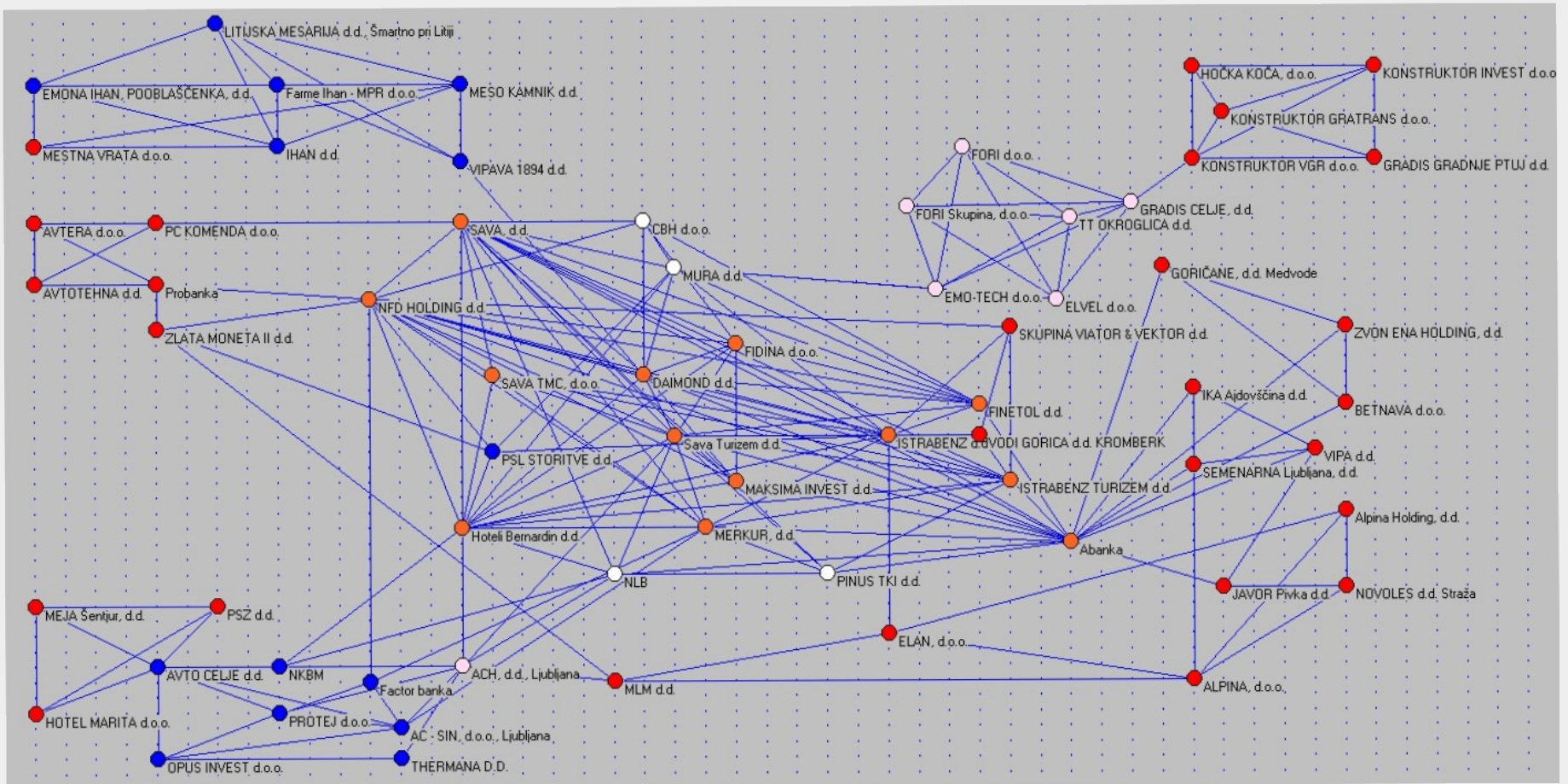
The Institute of Criminology at the Faculty of Law has obtained access to police records of all committed offenses...

Database contains description of committed offense, geolocation data, description of involved persons, description of method of offense, ID of persons, etc.

The incidence of crime

Data were imported into the database and simple OLAP analytic has been programmed…
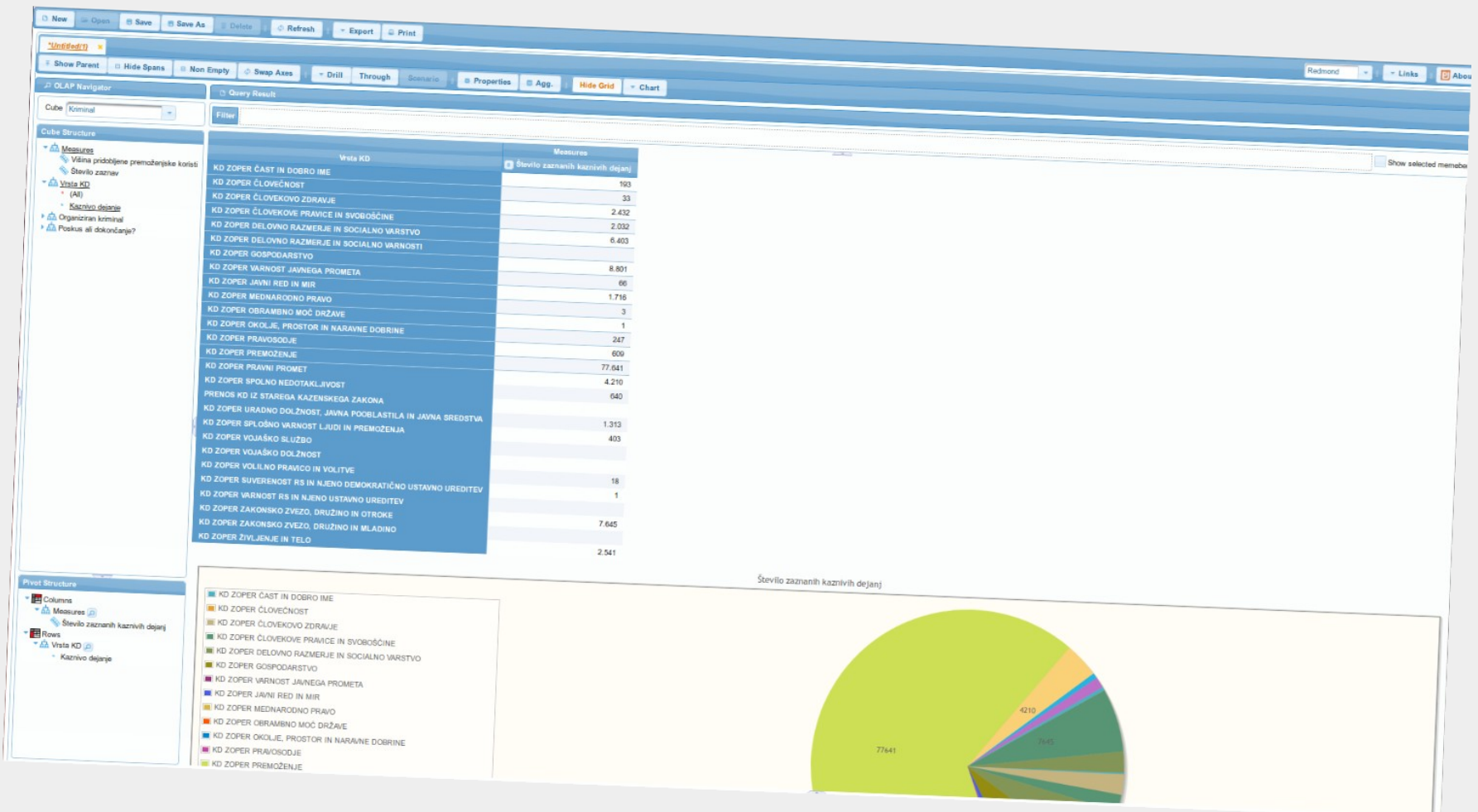
The incidence of crime

Data could also be easily imported in PSPP (software for advanced statistical analysis).

The incidence of crime

Some other possible analysis:

- visualisation of crime locations;
- time series;
- analysis of recidivism;
- clustering of criminal offences;
- crime predictions;
- ...

Data about lobbists and lobying contacts

We have received data about registered lobbying contacts from the *Commission for the Prevention of Corruption*. Unfortunately, data were not provided in machine readable form...

Data about lobbists and lobying contacts

... however, volunteers entered the data into the database...

Data about lobbists and lobying contacts

… Database was linked to business registry and some other databases and some analysis of lobbyists and their companies has been prepared.
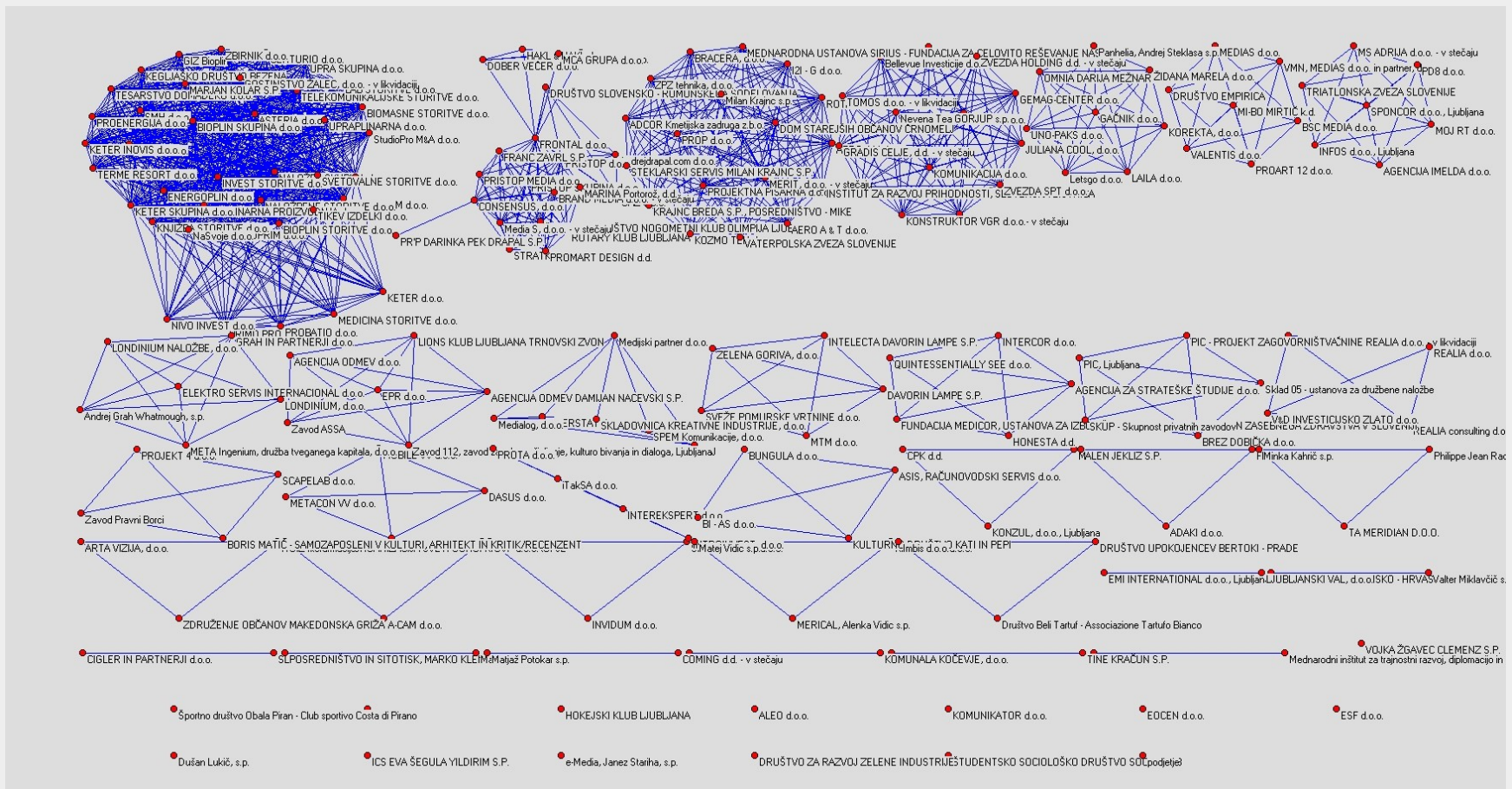
Data about lobbists and lobying contacts

Finally, visualisation of lobbying contacts has been created.

http://kdovpliva.si/
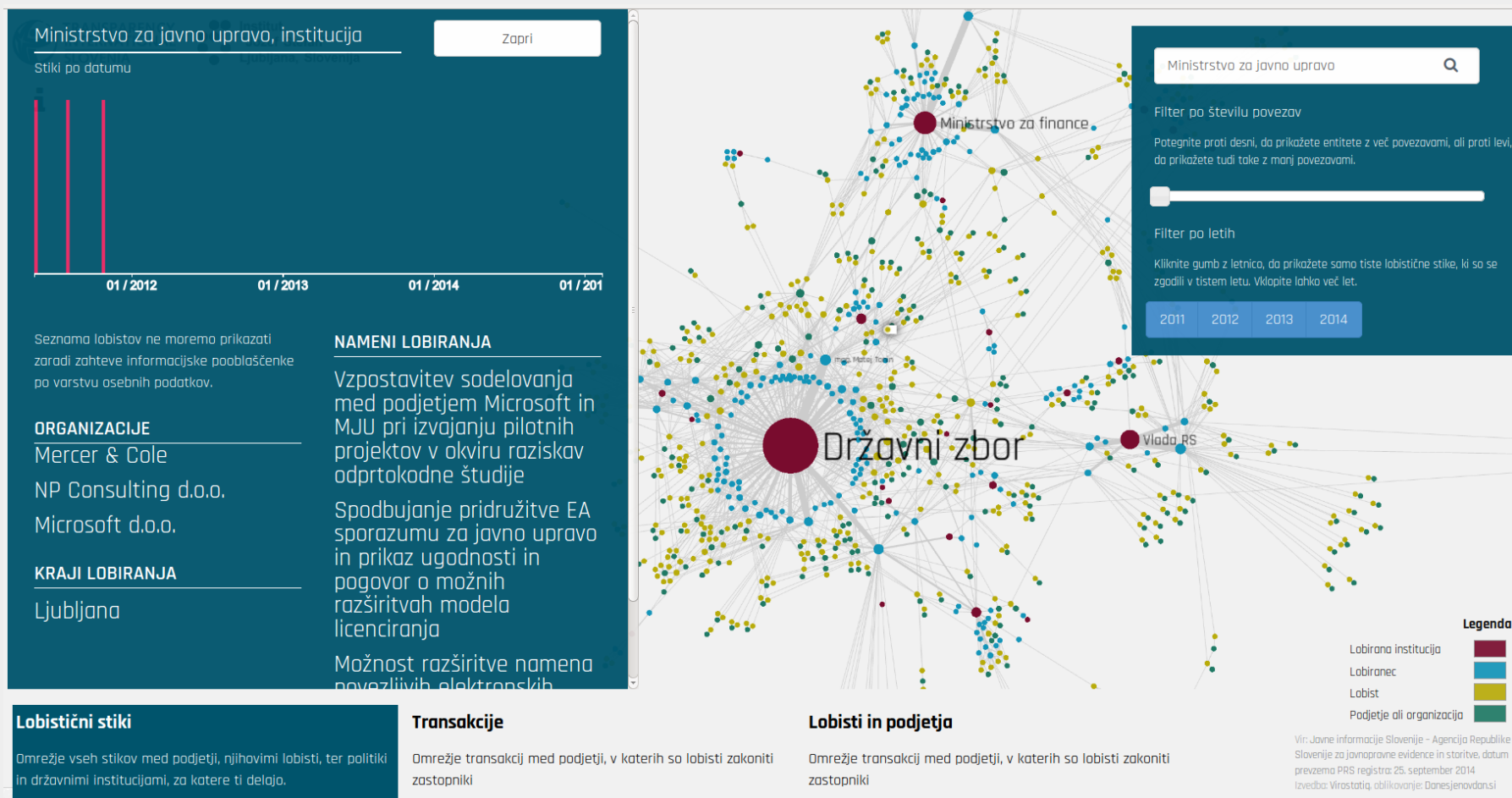
Legislative Monitor (zakonodajni-monitor.si)

Legislative Monitor is (was) an interactive tool which gives an insight into the legislation process, MEP's activities, analysis of their discussions, reported lobbying contacts, etc.
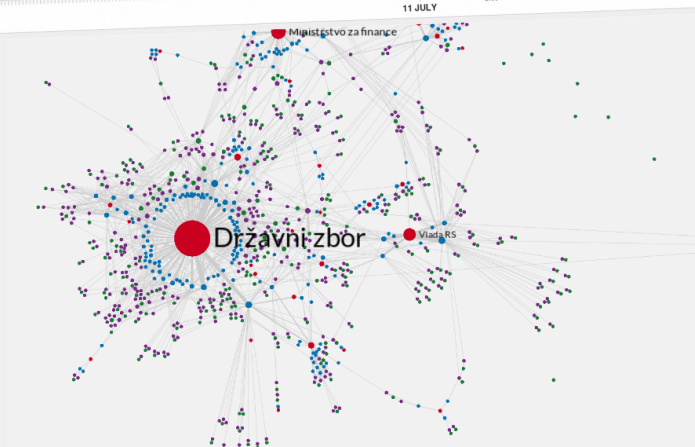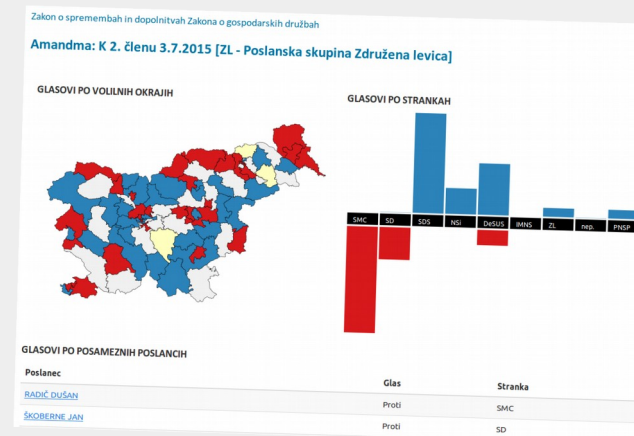
http://zakonodajni-monitor.si/
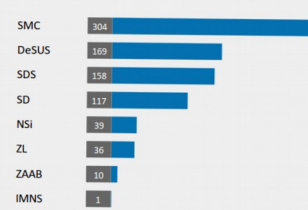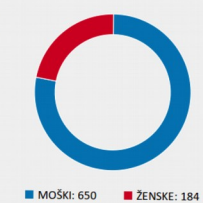
Salaries in public sector

Ajpes is collecting all data about salaries of civil servants. Through access of public information act we received the database of salaries from 2010. Database also contains (anonymised) ID's of civil servants.

Salaries in public sector

From the salaries database career paths could be computed...

Salaries in public sector

... and from career paths migrations of workforce among budget users could be computed and visualised.
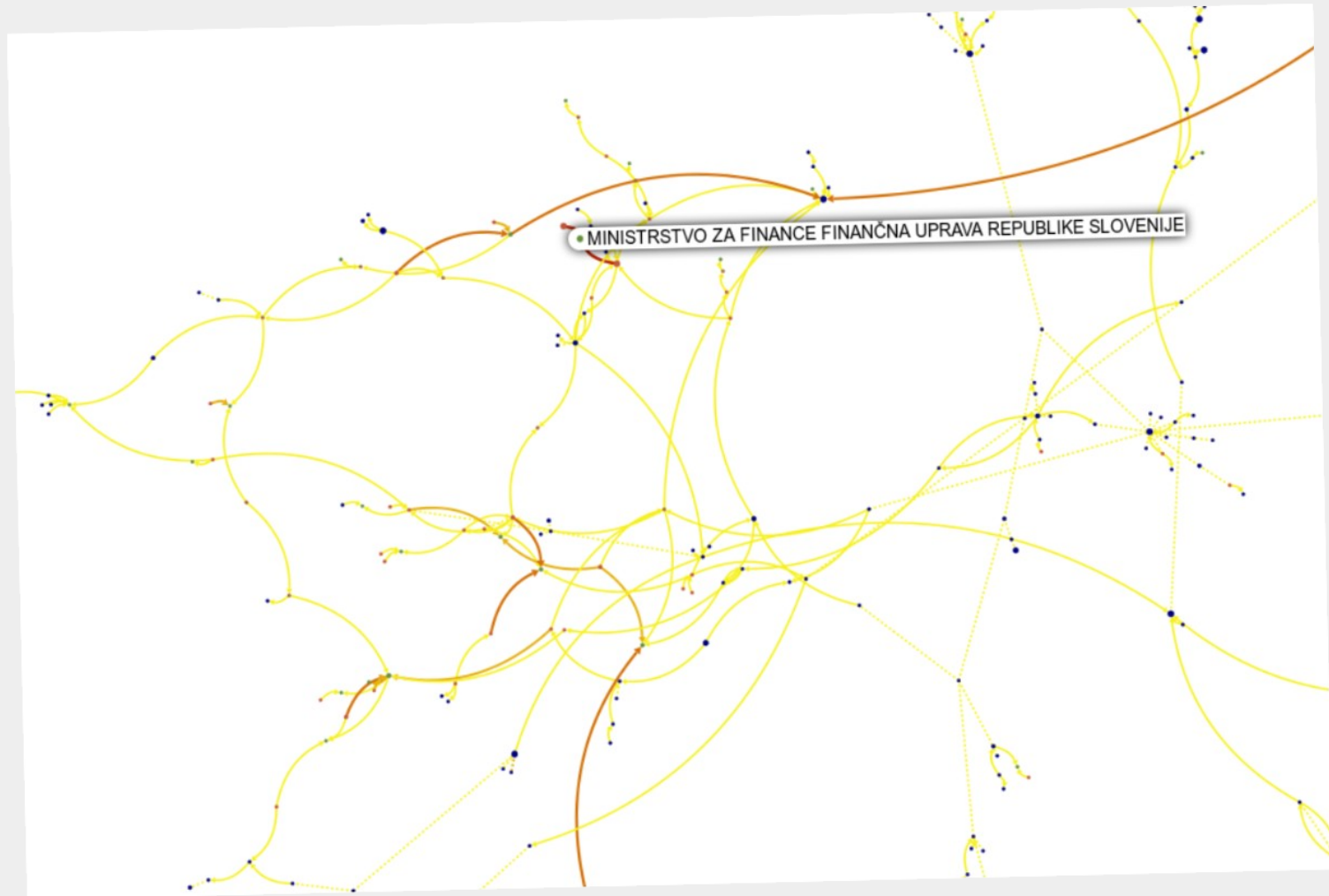
Clustering of employees migrations among public sector bodies.
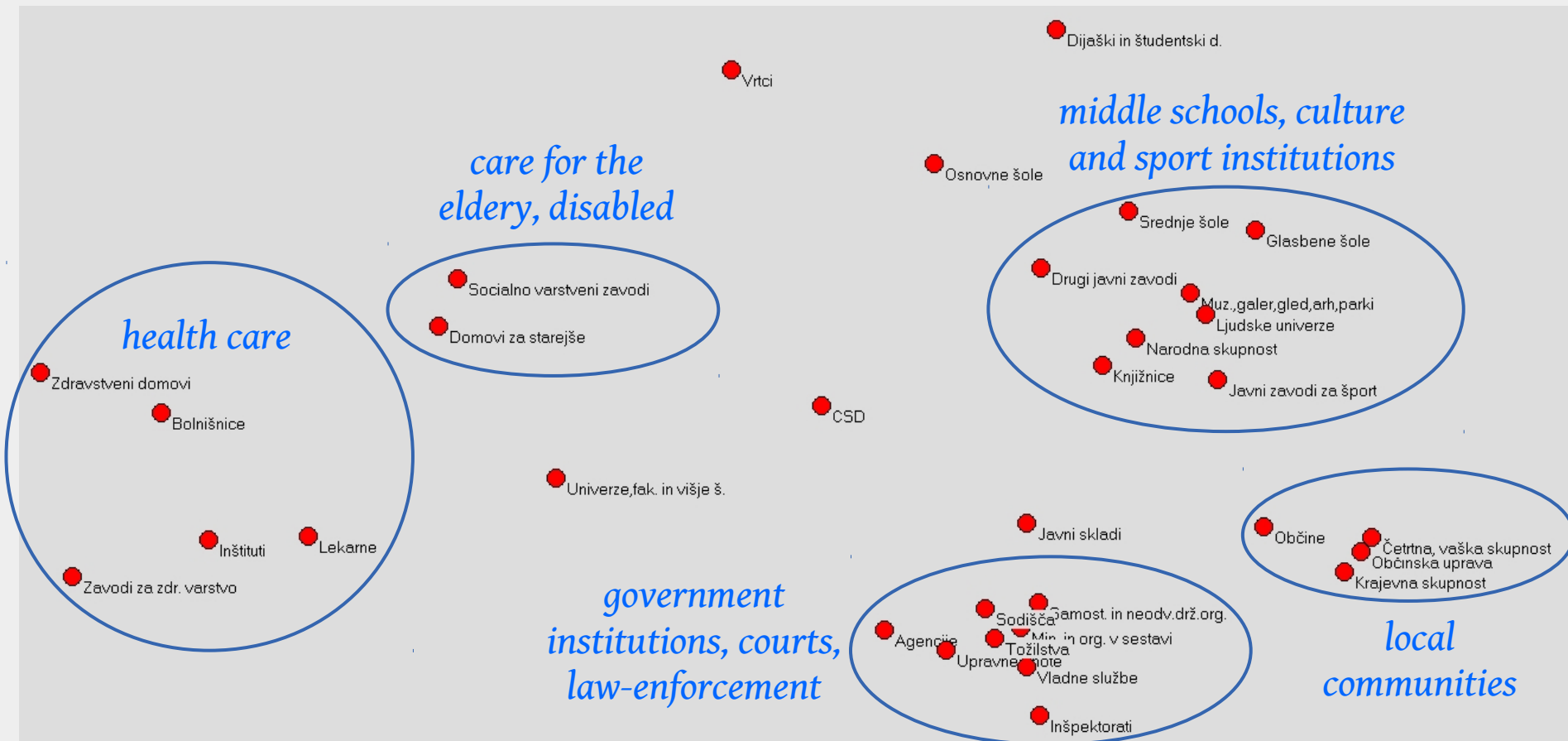
Another possible analysis:

- the influence of leadership change on staff recruitment;
- the influence of leadership change in sick leaves and terminations of employment contracts (possible indicator of mobbing);
- rate between minimum and maximum salary in public different parts of sector;
- amount of women in different employment positions;
- visualisations of wage premiums;
- what-if scenarios for salaries reform;
- ...

Public procurement database

We have received database of all Slovenian public procurements including small value procurements (published on a portal of public procurements).

Public procurement database

Some possible analysis:

- the influence of leadership change on public procurement;
- price comparison;
- visualisation of geographical distribution of granted procurements ("*Do the local companies usually win?*")
- detection of hidden consortium;
- ...

Visualisation of financial data

In 2016 we did a two small visualizations of open data: municipalities' budgets and development programs of Slovenian government.

http://www.transparency.si/projekti/drzavni-projekti/

http://www.transparency.si/projekti/proracuni-obcin/
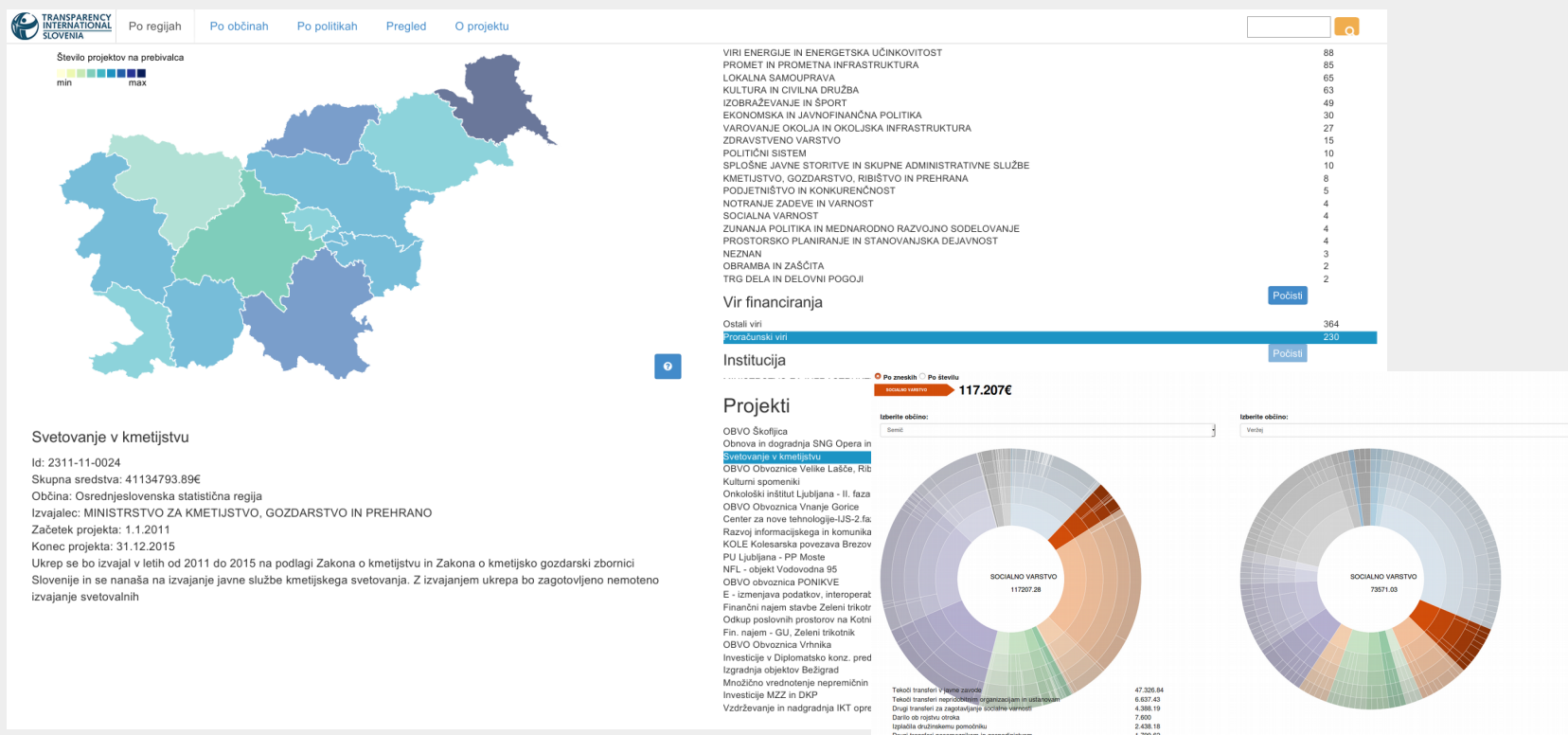
Financial transactions of public sector

- Supervizor (now Erar) is an online application which enables simple browsing through the financial transactions and their graphical presentation.
- Application was developed by *Commission for the Prevention of Corruption* in 2011. Later was renamed to Erar.

- User can view all money transfers from selected budget user or all money transfers from the budget user to selected company.

- Application also shows data about public procurements, information about the business entities in Slovenia, accounting entries of payments (for direct budget users), and some other additional data (e.g.: data about tax debtors and transfers to the so called favourable tax environments ("tax havens"), etc.

Financial transactions of public sector

Data on monthly sums of funds received from individu[al users]
of public funds
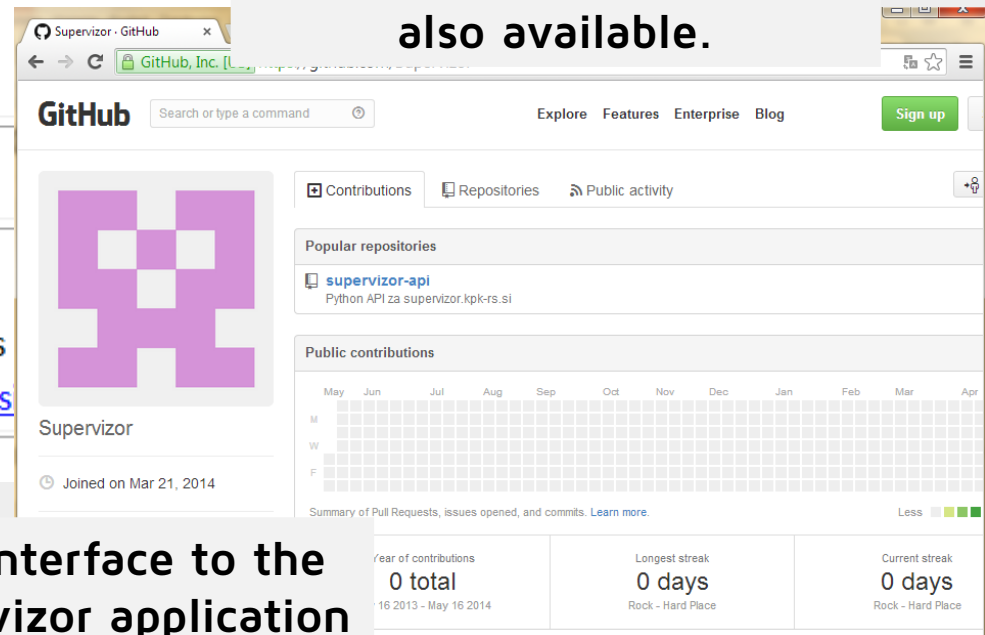
The data is located in a CSV text files (individual fields ar[e ...]
format with Unix/Linux end of the line. The file has a [...]
individual files contain data for each year (since 2003). F[or ...]
basis. Each file is compressed in. gz format.

Data structure description

field	field description
sifra_pu (*budget_user_code*)	Budget user code. Codes of budget users (http://www.ujp.gov.s[i ...]

Part of Erar data is available for download in machine readable form.

Description of data ~~is~~ was also available.

API interface to the Supervizor application ~~is~~ was also provided.

http://erar.kpk-rs.si

Financial transactions of public sector

With Erar database of financial transactions of public sector several analysis is possible. For instance:

- analysis of payments to media companies;
- analysis of grants paid to private companies;
- correlation of government in charge with financial flows;
- detecting interesting financial flows patterns;
- ...

Financial transactions of public sector

Grants to media companies

Grants from January, 1st 2003 till October, 1st 2014
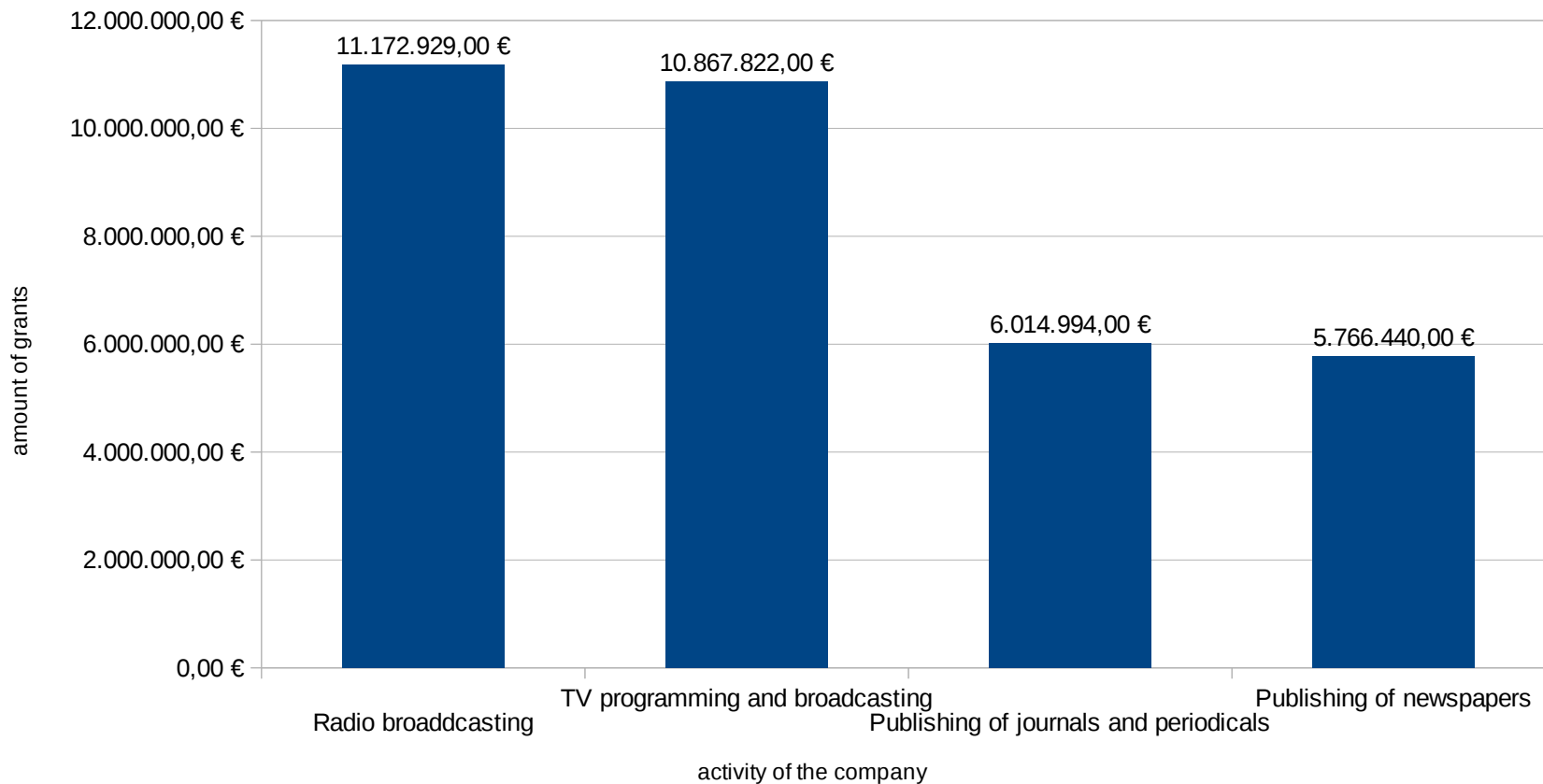
Amount of grants from Slovenian budget users to media companies. Media companies were identified from business registry by their main activity registered.

All financial transactions to media companies

Payments from January, 1st 2003 till October, 1st 2014.
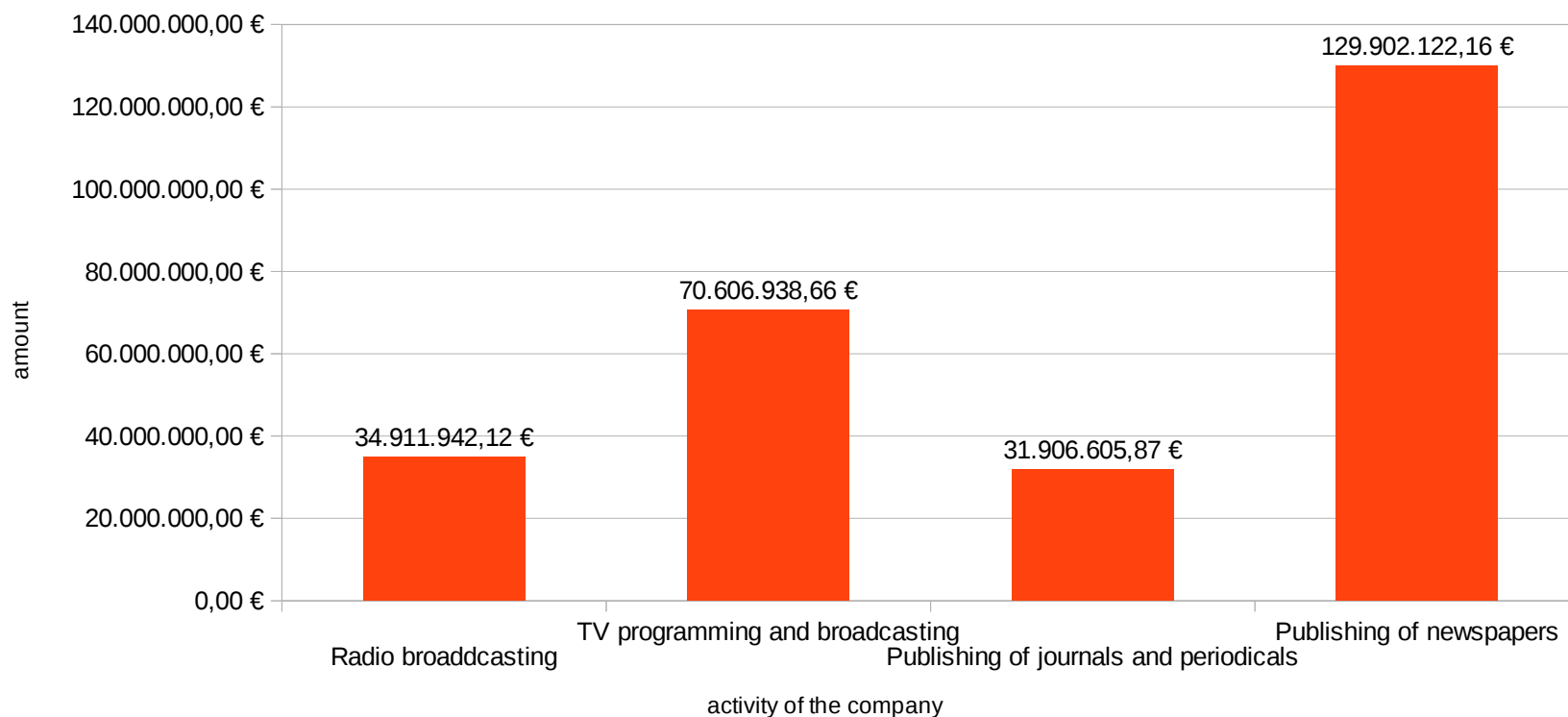
Amount of payments from Slovenian budget users to media companies. Media companies were identified from business registry by their main activity registered.

Annual transactions to companies registered to...

Legend:
- Public relation and communication activities
- Media representation
- Market research and public opinion polling
- Advertising agencies
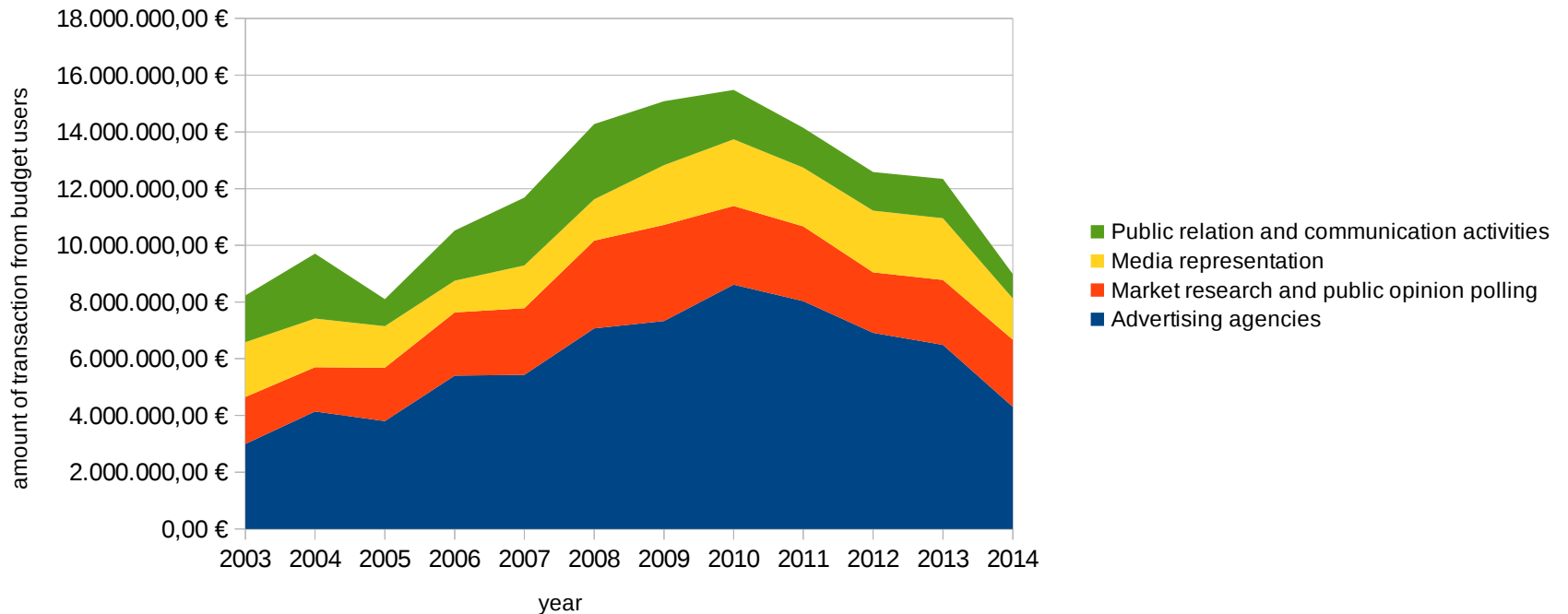
Annual amount of payments from Slovenian budget users to companies registered for management consultancy, PR, advertising, marketing research, etc. Types of companies were identified from business registry by their main activity registered.

Financial transactions of public sector

Občina: ŽALEC

Znesek: 0.2%
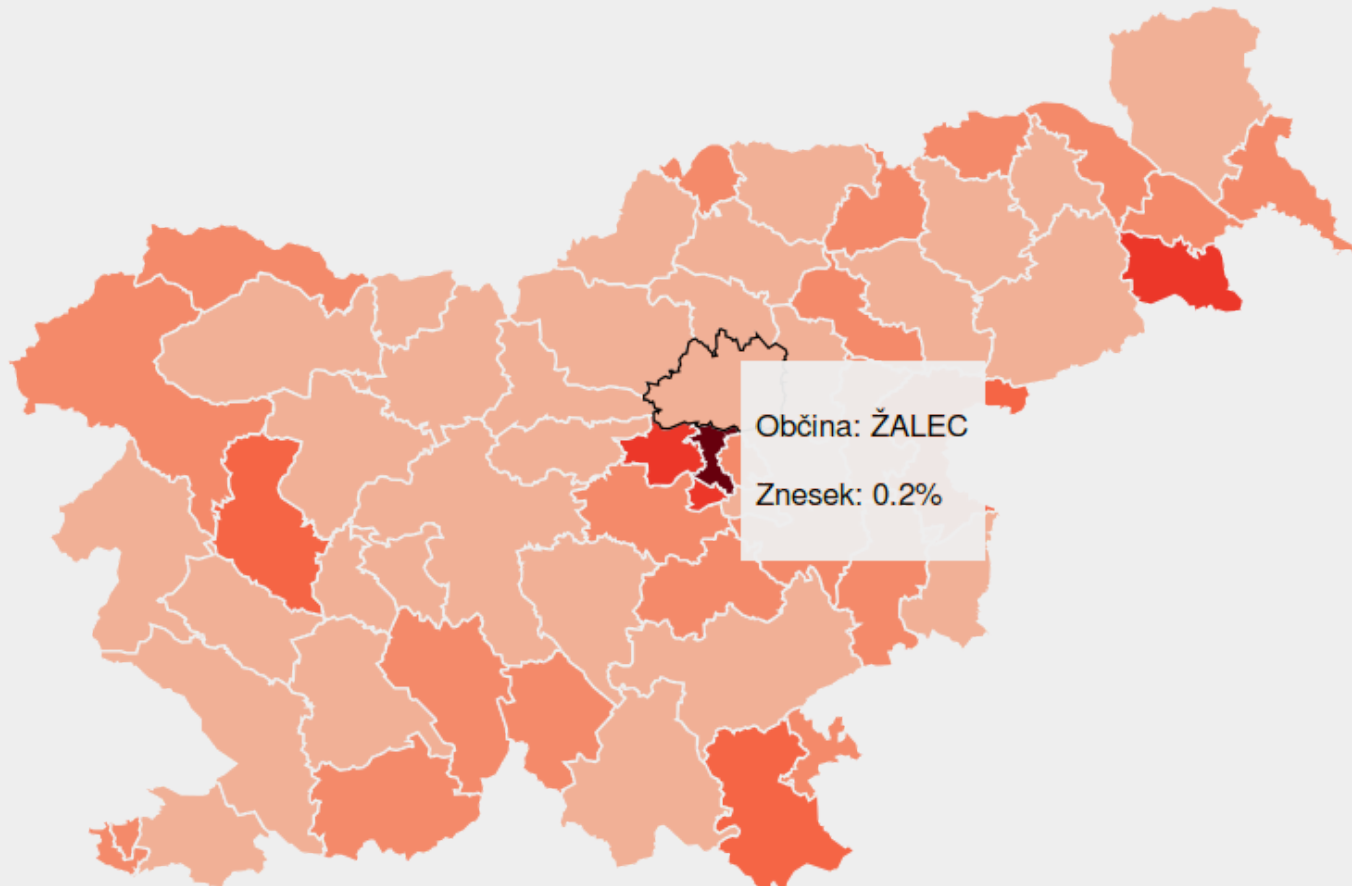
Visualisation of share of grants paid to companies
located in different municipalities.

Financial flow analysis

Summary of monthly payments to 65 companies

Change of government
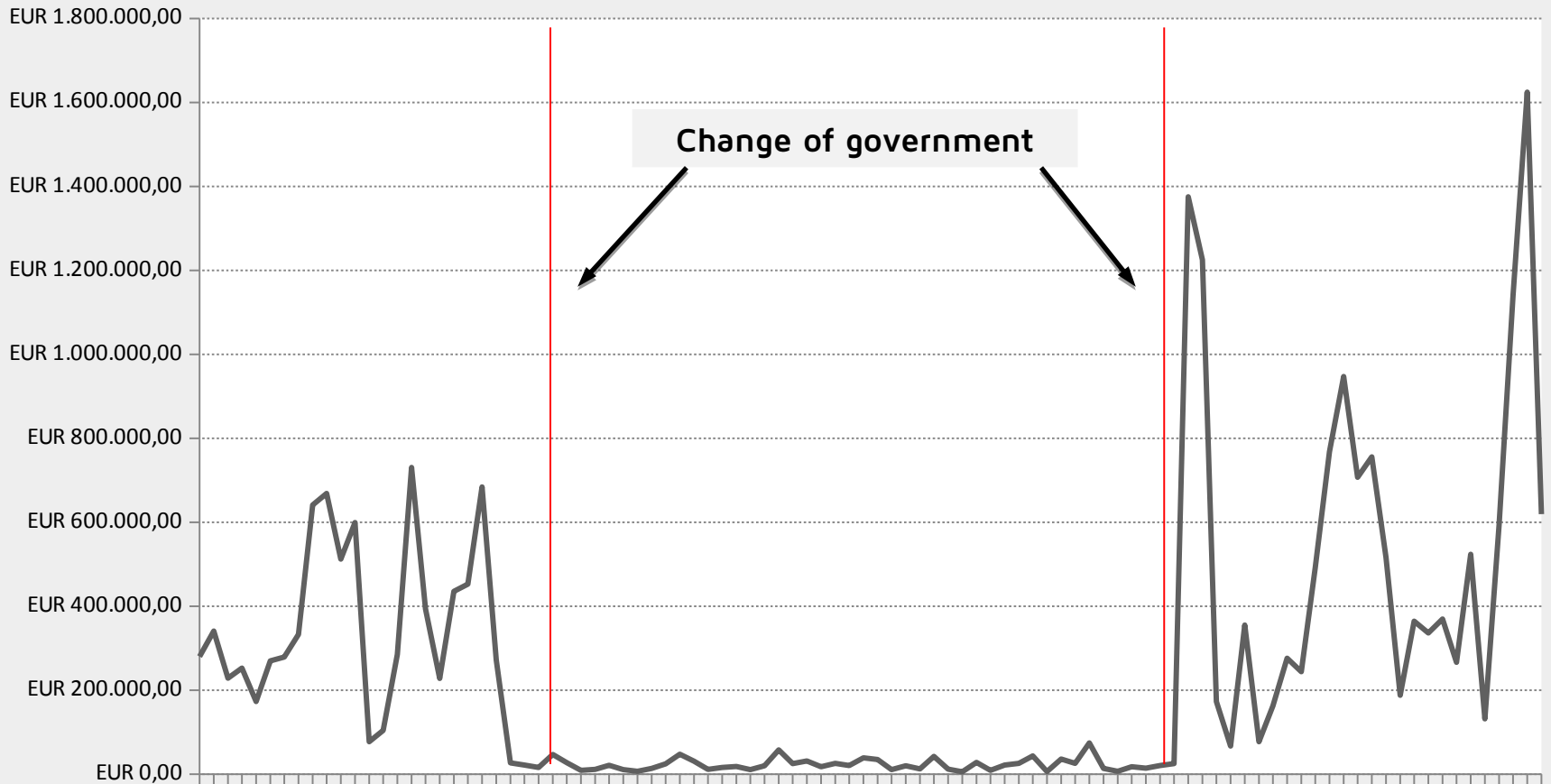
Financial flow analysis

Summary monthly payments to 252 companies
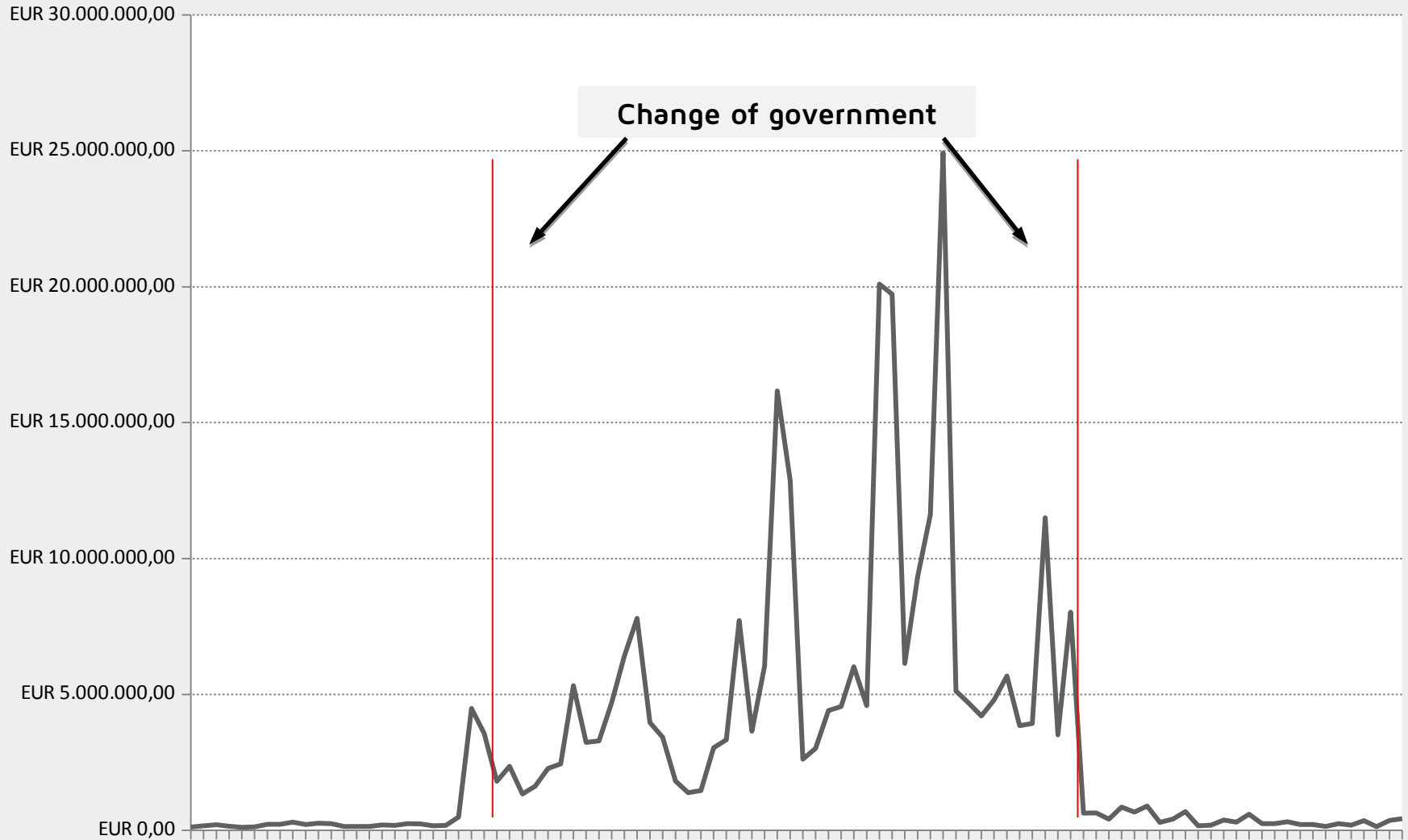
Monthly payments to companies registerd for 'IT activities'

In Slovenian Business Register there are 6223 companies registered for 'IT activities', 3796 of them had business with budget users.

Some other datasets:

- impositions of sanctions for offences and inspector decisions;
- list of fax debtors (published on *Tax Administration's* website);
- database of financial transfers to the so-called favourable tax environments ("tax havens"), published on the *Office for Money Laundering Prevention* website;
- accounting entries of payments for direct budget users from MFERAC database (is already included in Erar database, but not published in machine readable form yet);

Other Supervizor/Erar datasets

Some other datasets:

- list of restrictions of business due to conflict of interest, published by *Commission for Prevention of Corruption*;
- database of fiscal stimulus – "*Government for the development of Slovenia*";
- financial transactions from European founds;
- adopted legislation;
- election candidates;
- list of registered media companies;
- ...

Other Supervizor/Erar datasets

Visualisation of tax debtors (author Gašper Žejn,
<http://www.kiberpipa.org/~hruske/stuff/durs.html>).

Example: Detection of restriction on business activities violations

Restrictions on business activities

Article 35 of the Integrity and Prevention of Corruption Act
(Restrictions on business activities and the consequences of violations)

(1) A public sector body or organisa1on which is committed to conducting a public procurement procedure in accordance with the regulations on public procurement or which carries out the procedure for granting concessions or other forms of public-private partnership, may not order goods, services or construction works, enter into public-private partnerships or grant special and exclusive rights to entities in which the official who holds office in the body or organization concerned or in cases where the official's family member has the following role:
– participating as a manager, management member or legal representative; or
– has more than a 5% level of participation in the founders' rights, management or capital, either by direct participation or through the participation of other legal persons.

Article 35 of the Integrity and Prevention of Corruption Act

(Restrictions on business activities and the consequences of violations)

...

(5) Officials shall communicate the name, registration number and head office of those entities with which they or their family members have a relationship, as specified in paragraph ti of this Article, to the body in which they hold office within one month after taking office and then no later than within eight days of any change occurring. The body shall submit the list of entities referred to in the preceding sentence of this Article to the Commission no later than within ti5 days of receipt of information on or a notification of changes regarding the entities. The Commission shall publish the list of entities referred to in the first sentence of this paragraph on its website every month.

...

List of restrictions

Supervizor/Erar database in machine readable form

Overview of reported restrictions

Integrity And Prevention of Corruption Act was enacted in 2003

Overview of reported restrictions

Overview of reported restrictions


```
matej@cryptoloop: ~
OBČINA ODRANCI                          | OSNOVNA ŠOLA ODRANCI
MESTNA OBČINA VELENJE                   | OSNOVNA ŠOLA ŠALEK VELENJE
OBČINA ŠTORE                            | OSNOVNA ŠOLA ŠTORE - VRTEC LIPA ŠTORE
MESTNA OBČINA KOPER                     | POKRAJINSKI MUZEJ KOPER, MUSEO REGIONALE DI CA
OKROŽNO SODIŠČE V MARIBORU              | RADIOTELEVIZIJA SLOVENIJA javni zavod, Ljublja
MESTNA OBČINA NOVO MESTO                | RAZVOJNI CENTER NOVO MESTO, Svetovanje in razv
OBČINA ČRENŠOVCI                        | SAUBERMACHER - KOMUNALA - podjetje za
OBČINA ŠKOFJA LOKA                      | SMUČARSKO TURISTIČNI CENTER STC STAR
OBČINA IVANČNA GORICA                   | SREDNJA ŠOLA JOSIPA JURČIČA IVANČNA
MESTNA OBČINA MARIBOR                   | ŠPORTNI OBJEKTI MARIBOR
OBČINA LENDAVA - LENDVA KÖZSEG          | TURIZEM LENDAVA d.o.o.
OBČINA PODČETRTEK                       | TURIZEM PODČETRTEK, gospodarsko in
MESTNA OBČINA VELENJE                   | VETERINA ŠOŠTANJ d.o.o., Šoštanj,
OBČINA ROGATEC                          |            ZAVOD OSNOVNA
OBČINA KOČEVJE
OBČINA ŠKOFJA LOKA
OBČINA LENDAVA - LENDVA KÖZSEG |
OBČINA SLOVENSKE KONJICE
OBČINA ŠMARJE PRI JELŠAH
MESTNA OBČINA VELENJE
OBČINA ŽALEC
(42 rows)

(END)
```

ePRS - POSLOVNI REGISTER SLOVENIJE

Preverite boniteto eS.BON
Več v FI-PO

Podatki o subjektu TURIZEM LENDAVA d.o.o., matična številka: 2080036000

osnovni podatki o subjektu	družbeniki in poslovni deleži	osebe, pooblaščene za zastopanje	člani organa nadzora	skupščinski sklepi	razno	sprememba družbene pogodbe / statuta

DRUŽBENIKI

zap. št. družbenika: 492897
identifikacijska številka: 5874645000 - matična številka (pravna oseba je vpisana v PRS)
firma: OBČINA LENDAVA - LENDVA KÖZSEG
Naslov: Lendava, Glavna ulica 20, 9220 Lendava - Lendva
vrsta odgovornosti za obveznosti družbe: ne odgovarja
datum vstopa: 27.01.2005

zap. št. družbenika: 576081
identifikacijska številka: 2220385000 - matična številka (pravna oseba je vpisana v PRS)
firma: ZVEZA DRUŠTEV ZA MEDKULTURNI DIALOG MURANIA - MURANIA-EGYESÜLETEK SZÖVETÉGE A KULTÚRÁK KÖZÖTTI PÁRBESZÉDÉRT
Naslov: Lendava, Bánffyjev trg 1, 9220 Lendava - Lendva
vrsta odgovornosti za obveznosti družbe: ne odgovarja
obseg odgovornosti: /
datum vstopa: 20.01.2009

zap. št. družbenika: 740816
identifikacijska številka: 5301971000 - matična številka (pravna oseba je vpisana v PRS)

Company established by the municipality

Cleaning the data

For data cleaning we need *Registry of budget users* (available on web) and *Business registry of Slovenia* (accessible through Access to Public Information Act)…

In first step we…

- remove entities which are in fact budget users;
- remove business entities established by budget users;
- *for other business entities add type, for associations we add type of association*;
- remove »*associations working in public interest*« (firemen associations, hunting and fishing associations, handicapped persons associations);
- remove political parties.

Cleaning the data

```
omejitve_poslovanja.sql (~/Namizje/datoteke/open_data/omejitve_poslovanja) - gedit
Odpri    Shrani    Razveljavi

omejitve_poslovanja.sql ×    Supervizor_psql.txt ×

# zacasna tabela
select organ, kpk_omejitve_poslovanja.sifra_pu, tip_omejitve, ps_naziv, ps_davcna, datum_od, datum_do,
datum_transakcije, znesek_transakcije, namen into temp_krsitve_omejitev from kpk_omejitve_poslovanja,
supervizor_transakcije where (kpk_omejitve_poslovanja.sifra_pu = supervizor_transakcije.sifra_pu) and
(kpk_omejitve_poslovanja.ps_davcna = supervizor_transakcije.davcna_stevilka::text) and (datum_od <
datum_transakcije) and (datum_transakcije < datum_do) and (extract(year from datum_transakcije) >= 2014);

# ali je prejemnik PU?
alter table temp_krsitve_omejitev add column ps_je_pu boolean;
update temp_krsitve_omejitev set ps_je_pu = FALSE;
update temp_krsitve_omejitev set ps_je_pu = TRUE from rpu where (ps_davcna = rpu.davcna::text);

# tip prejemnika
alter table temp_krsitve_omejitev add column ps_oblika text;
update temp_krsitve_omejitev set ps_oblika = kategorija from rpu where (ps_je_pu = TRUE) and (ps_davcna =
rpu.davcna::text);
update temp_krsitve_omejitev set ps_oblika = oblika from prs_enota_rs where (ps_je_pu = FALSE) and
(ps_davcna = prs_enota_rs.davcna::text);
update temp_krsitve_omejitev set ps_oblika = drustvo from prs_enota_rs where (ps_je_pu = FALSE) and
(ps_oblika ~* 'društvo') and (ps_davcna = prs_enota_rs.davcna::text);
update temp_krsitve_omejitev set ps_oblika = trim(ps_oblika);

# ali je ustanovitelj prejemnika PU
alter table temp_krsitve_omejitev add column ps_ustanovil_pu boolean;
update temp_krsitve_omejitev set ps_ustanovil_pu = FALSE;
update temp_krsitve_omejitev set ps_ustanovil_pu = TRUE from prs_enota_rs, prs_ustanovitelji, rpu where
(ps_davcna = prs_enota_rs.davcna::text) and (prs_enota_rs.enota = prs_ustanovitelji.enota) and
(p_subjekt_maticna = rpu.maticna);

SQL ▾    Širina _tabulatorja: 8 ▾    Vr. 52, St. 275    VST
```

Results

- Number of all public bodies reported restrictions of business: **294**
- Number of all public bodies, where financial transactions were found in 2014 and 2015: **154**
- Number of public bodies where there is suspicion of violation (cleaned data): **148**

- Number of suspicious transactions in 2014: **2.912.193,03 EUR**
- Number of suspicious transactions till end of April 2015: **1.658.190,47 EUR**

- **Finally we have 391 pairs public body – company on the list**. Can we remove more data?
  - all remaining associations and clubs?
  - all companies that contain word »*public*« in their nave, even if they are not registered as public companies?

Additional cleaning of data

```
 prejemki   | st_transakcij |           organ                    |
------------+---------------+------------------------------------+--------------------------------------------------
  63744.8000 |          734 | MESTNA OBČINA MARIBOR               | MARIBORSKI VODOVOD javno podjetje d.d.
 111916.4800 |          110 | OKROŽNO SODIŠČE V KOPRU             | BANKA KOPER d.d. Dvojezična firma: BANKA KOPER d.d. - s.p.a
   5972.1400 |           75 | MESTNA OBČINA NOVA GORICA           | GEN-I, trgovanje in prodaja električne energije, d.o.o.
1107889.0400 |           72 | MESTNA OBČINA LJUBLJANA             | WALDORFSKA ŠOLA LJUBLJANA
  34322.9900 |           62 | OBČINA BRDA                         | VINSKA KLET "GORIŠKA BRDA" druga, kmetijstvo, proizvodnja
 140529.6400 |           33 | MESTNA OBČINA KRANJ                 | ZAVOD CARNICA - Zavod za kulturo in turizem
  16074.8600 |           32 | OBČINA POSTOJNA                     | ZVEZA DRUŠTEV MLADINSKI CENTER POSTOJNA
   8815.0800 |           30 | OKRAJNO SODIŠČE V LJUBLJANI         | Odvetniška pisarna Lebar Miklič, Rozman d.o.o.
   6291.4900 |           30 | OBČINA SLOVENSKA BISTRICA           | KADRING - kadrovsko in poslovno svetovanje, d.o.o.
   6166.9200 |           30 | OBČINA PIVKA                        | NEMEZA, finančni inženiring in storitve d.o.o.
  12580.6100 |           28 | OBČINA VOJNIK                       | MPRINT servis in prodaja računalniške opreme, d.o.o.
  10139.1800 |           28 | MESTNA OBČINA MURSKA SOBOTA         | DRUŠTVO UPOKOJENCEV MURSKA SOBOTA
  12073.6700 |           25 | OKROŽNO SODIŠČE V LJUBLJANI         | OREL LADO - ODVET
  35952.0200 |           24 | MESTNA OBČINA MARIBOR               | DRUŠTVO ZA POMOČ SAMOPOMOČ NA PODROČJU ZASVOJENOSTI ZDRA
  23404.4200 |           24 | OBČINA ČRNOMELJ                     | MLADINSKI CENTER BIT angleški prevod firme: YOUTH CENTER BI
  10809.2000 |           24 | OBČINA SLOVENSKE KONJICE            | EKOTIP - KEROVAN KOLAR proizvodnja in storitve, d.n.o. S
  37560.1800 |           23 | OBČINA ŽIRI                         | M SORA, trgovina in proizvodnja, d.d.
  15747.6800 |           23 | OBČINA SLOVENSKA BISTRICA           | KULTURNO IZOBRAŽEVALNO DRUŠTVO KERAMIKOV IN LONČARJEV PODRA
    914.4200 |
  17507.1400 |
   6997.4500 |
  24742.0000 |
   3386.7200 |
   4520.0100 |
  55780.5400 |
  25932.2500 |
   4958.6300 |
   2160.7300 |
  44870.0000 |
  38306.9600 |
  13881.8400 |
   6866.1600 |
   4829.0400 |
   4000.0000 |
  41217.0000 |
   5527.1300 |
  40160.1100 |
   4301.9300 |
   4080.2500 |
```

Podatki o subjektu MARIBORSKI VODOVOD javno podjetje d.d., matična številka: 5067880000

Preverite boniteto eS.BON
Več v FI-PO

| osnovni podatki o subjektu | družbeniki in poslovni deleži | osebe, pooblaščene za zastopanje | člani organa nadzora | skupščinski sklepi | razno | sprememba družbene pogodbe / statuta |

OSNOVNI PODATKI O SUBJEKTU

status subjekta:	vpisan
datum vpisa subjekta v sodni register:	23.04.1991
matična številka:	5067880000
davčna številka:	SI 68041527
firma:	MARIBORSKI VODOVOD javno podjetje d.d.
skrajšana firma:	MARIBORSKI VODOVOD d.d.
sedež:	Maribor
poslovni naslov:	Jadranska cesta 24, 2000 Maribor
pravnoorganizacijska oblika:	Delniška družba d.d.
število delnic:	285.489
vrsta organa nadzora:	Nadzorni svet

Type of company: Joint-stock company (d.d.)
However from the name of the company we can deduce that it is public company

Results

- Transactions of remained business entities:
  3.510.827,56 EUR.

Overview of purposes of transactions

```
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  431.8800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | GRAD, računalniški programi in inženiring d.d
  436.6000 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | TRGOVINA, POSREDNIŠTVO IN ZASTOPSTVO DRAFT MI
  441.7900 |                                | OKROŽNO SODIŠČE V KOPRU          | OREL LADO - ODVETNIK
  447.7900 |                                | OKROŽNO SODIŠČE V MARIBORU       | BANKA KOPER d.d. Dvojezična firma: BANKA KOPE
  451.4000 |                                |                        LJANI     | ŠOŠTERIČ FRANCI - ODVETNIK
  478.8500 |                                |                                  | Odvetniška družba Fašun, Melihen, Milač, Stro
  494.9000 |                                |                                  | ISTRABENZ PLINI, plini in plinske tehnologije
  531.0200 |                                |                        LJANI     | Odvetniška družba Fašun, Melihen, Milač, Stro
  554.6100 |                                |                        LJANI     | OREL LADO - ODVETNIK
  609.0000 |                                |                        LJANI     | Odvetniška družba Fašun, Melihen, Milač, Stro
  632.5700 |                                |                        LJANI     | Odvetniška družba Fašun, Melihen, Milač, Stro
  638.7500 |                                |                        RU        | BANKA KOPER d.d. Dvojezična firma: BANKA KOPE
  643.5500 |                                |                        LJANI     | Odvetniška pisarna Leben, Miklič, Rozman d.o.
  653.9200 |                                |                        LJANI     | OREL LADO - ODVETNIK
  707.4800 |                                |                        LJANI     | OREL LADO - ODVETNIK
  772.7700 |                                | OKROŽNO SODIŠČE V MARIBORU       | ŠOŠTERIČ FRANCI - ODVETNIK
  799.5000 |                                | OKROŽNO SODIŠČE V KOPRU          | BANKA KOPER d.d. Dvojezična firma: BANKA KOPE
  800.0000 |                                | OKROŽNO SODIŠČE V KOPRU          | BANKA KOPER d.d. Dvojezična firma: BANKA KOPE
  847.9000 |                                | OKROŽNO SODIŠČE V MARIBORU       | ŠOŠTERIČ FRANCI - ODVETNIK
  854.2400 |                                | OKRAJNO SODIŠČE V LJUBLJANI      | Odvetniška pisarna Leben, Miklič, Rozman d.o.
  866.0800 |                                | OKRAJNO SODIŠČE V LJUBLJANI      | OREL LADO - ODVETNIK
  868.0300 |                                | OKRAJNO SODIŠČE V LJUBLJANI      | Odvetniška pisarna Leben, Miklič, Rozman d.o.
 1207.6800 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | OREL LADO - ODVETNIK
 1212.4700 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | MERKUR - trgovina in storitve, d.d.
 1591.0000 |                                | OKROŽNO SODIŠČE V LJUBLJANI      | OREL LADO - ODVETNIK
 3956.0000 | VL 2011/187093 VRAČILO VARŠČINE | Okrajno sodišče v Šmarju pri Jelšah | JAGROS trgovina, proizvodnja in storitve, d.o
 4119.8000 | VL 2011/187093 VRAČILO VARŠČINE | Okrajno sodišče v Šmarju pri Jelšah | JAGROS trgovina, proizvodnja in storitve, d.o
 5482.4000 | 2986445 STORITVE-F.O.          | OKROŽNO SODIŠČE V MARIBORU       | ŠOŠTERIČ FRANCI - ODVETNIK
18037.5200 | VL 2011/187093 VRAČILO VARŠČINE | Okrajno sodišče v Šmarju pri Jelšah | JAGROS trgovina, proizvodnja in storitve, d.o
          | IN 222/12 DOLŽNIK MENTOR ČERIFOSKI | OKROŽNO SODIŠČE V KOPRU       | BANKA KOPER d.d. Dvojezična firma: BANKA KOPE
(245 rows)
```

Series of transactions with the same amount of money – it seems there is some kind of periodic payments. Payments of service?

»VRAČILO VARŠČINE« (RETURN OF JUDICIAL SECURITY) - probably there is no business.
»STORITVE F. O.« - services of natural persons?

Overview of purposes of transactions

- ŠOŠTANJ MUNCIPALITY -> NTS, navtika, trgovina in storitve, d.o.o.
  - 2014-09-30, 2.547,99 €
  - »*ASIGNACIJA*« (*ASIGNATION*)

- ŽIRI MUNCIPALITY -> M SORA, trgovina in proizvodnja, d.d.
  - 2015-02-23, 4.052,90 €
  - »*TAX REFUND*«

- ŽELEZNIKI MUNCIPALITY -> DOMEL, Elektromotorji in gospodinjski aparati, d.o.o.
  - 2014-07-30, 20.000,00 €
  - »*RN41412014042NAKUP ENOSOBNEGA*« (*THE PURCHASE OF ONE-ROOM (APARTMENT)*)

- SLOVENSKE KONJICE MUNCIPALITY -> MLADINSKI CENTER DRAVINJSKE DOLINE
  - 2014-11-18, 2.280,00 €
  - »*PF 2014/8000555 SOF.MLAD.PROGRAMA*« (*CO-FINANCING OF YOUTH PROGRAM*)

- LOVRENC NA POHORJU MUNCIPALITY -> PERGOLA, družba za gradbeništvo, posredovanje, trgovino in druge storitve, d.o.o.
  - 2014-07-21, 38.253,60 €
  - »*PLAČILO RAČUNA PF 2011/D000180 004*« (*BILL PAYMENT*)

- MESTNA OBČINA CELJE, CELJE -> NT&RC Podjetje za časopisno in radijsko dejavnost d.o.o. Celje
  - 2015-04-10, 3.008,55 €
  - »*PLAČ.RAČ.-OPRAV.STORITVE*« (*BILL PAYMENT – SERVICES PROVIDED*)

Overview of purposes of transactions

- BRDA MUNCIPALITY -> VINSKA KLET "GORIŠKA BRDA" zadruga, kmetijstvo, proizvodnja vina in trgovine z.o.o. Dobrovo
  - 2014-10-06, 3.518,84 €
  - »*PL.RAČ. ZA VINO*« (*BILL PAYMET FOR VINE*)

- KRŠKO MUNCIPALITY -> AVTOBUSNI PREVOZI BOŽIDAR BRAČUN S.P.
  - 2015-04-01, 2.787,60 €
  - »*PREVOZI ŠOL. OTROK 2-2015 2015/300*« (*TRANSPORTATION OF SCHOOL CHILDREN*)

- BLED MUNCIPALITY -> PROMOTIVA, ORGANIZACIJA DOGODKOV, ANTOLIN GREGA S.P.
  - 2014-11-13, 3.000,00 €
  - »*RN?101-2014-OTROŠKI ANIMACIJSKI NA*« (*CHILDRENS ENTERTAINMENT*)

Overview of purposes of transactions

- ŽALEC MUNCIPALITY -> DRUŠTVO SVETNIKOV SDS ŽALEC:
  - 2014-10-27: 622,60 €
  - 2014-09-22: 877,20 €
  - 2014-09-02: 877,20 €
  - 2014-07-11: 877,20 €
  - 2014-06-24: 877,20 €
  - 2014-05-29: 877,20 €
  - 2014-04-22: 877,20 €
  - 2014-03-18: 2.631,60 €

- OKROŽNO SODIŠČE V LJUBLJANI -> GRAD, računalniški programi in inženiring d.d.:
  - 2015-04-13: 417,24 €
  - 2015-03-19: 417,24 €
  - 2015-02-25: 417,24 €
  - 2015-01-12: 431,88 €
  - 2014-12-10: 431,88 €
  - 2014-12-01: 431,88 €
  - 2014-10-13: 431,88 €
  - 2014-09-15: 431,88 €
  - 2014-08-14: 431,88 €
  - 2014-07-14: 431,88 €
  - 2014-06-27: 431,88 €
  - 2014-05-15: 431,88 €
  - 2014-04-14: 431,88 €
  - 2014-03-17: 431,88 €
  - 2014-02-14: 431,88 €
  - 2014-01-15: 431,88 €

Example: Text analysis of Slovenian legislation

Text analysis of Slovenian legislation

- Slovenian legislation (full text) was obtained in HTML format and imported into the database with some additional metadata.
- Data were exported to "named-line documents format": first word is the name of the document (act), after space/tab there is a full text of the document till the newline:

```
\COPY (select replace(naslov_zakona, ' ', '_'),
regexp_replace(vsebina_zakona, E'[\\n\\r\\u2028]+', ' ', 'g' ) from
zakonodaja_zakoni, zakonodaja_besedila where (zakonodaja_zakoni.sop
= zakonodaja_besedila.sop) and (id ~* 'zak')) to
'slovenian_acts.txt' delimiter E'\t';
```

- Slovenian characters (čšž) were removed:

```
sed 's/\č/c/g' slovenian_acts.txt | sed 's/\š/s/g' | sed 's/\ž/z/g'
| sed 's/\Č/C/g' | sed 's/\Š/S/g' | sed 's/\Ž/z/g' >
slovenian_acts_csz.txt
```

- OntoGen application is quite old and unfortunately does not use UTF-8 encoding, so file needs to be encoded in CP1250:

```
iconv -c -f utf-8 -t cp1250 slovenian_acts_csz.txt >
slovenian_acts_win.txt
```
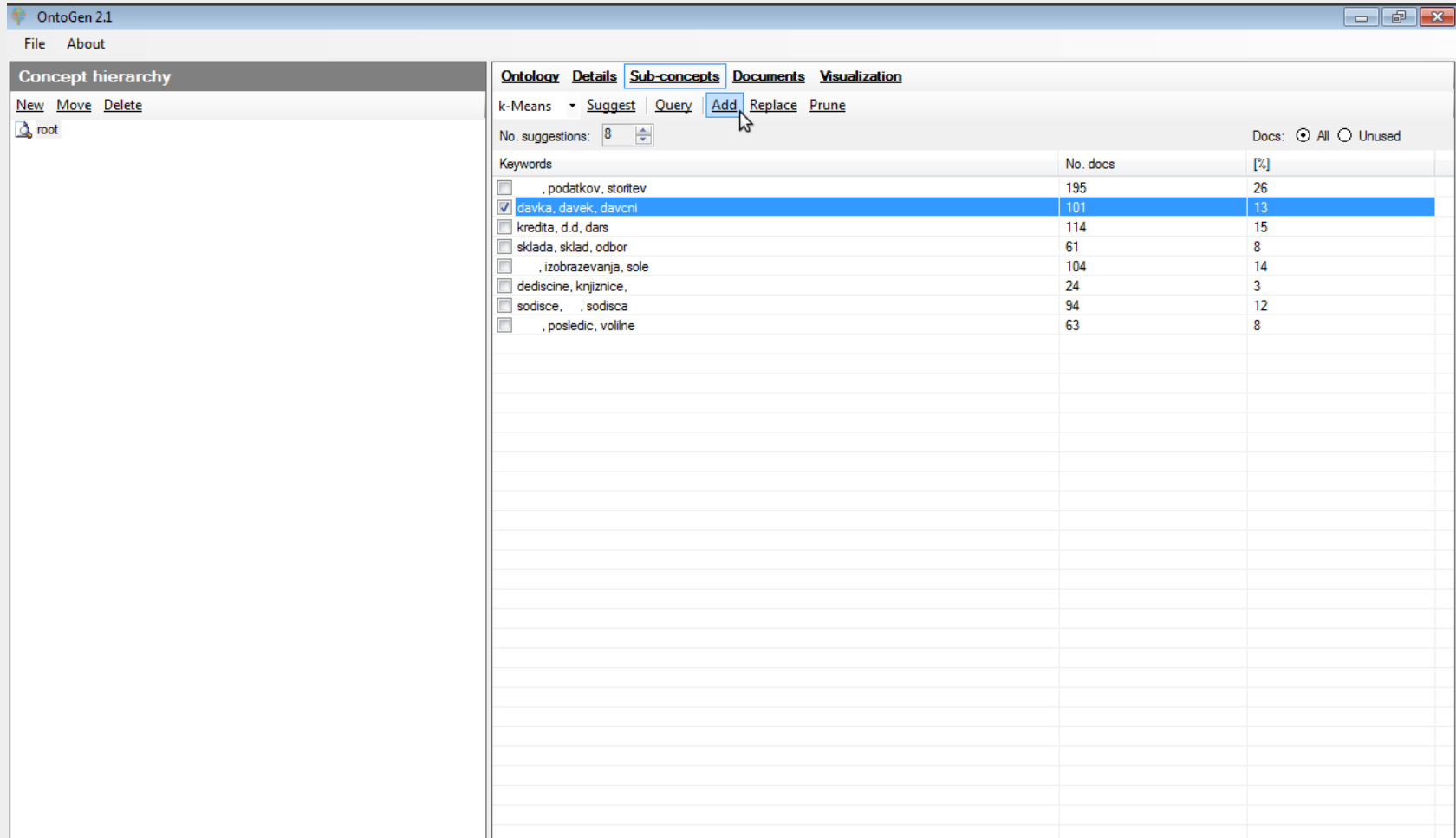
Text analysis of Slovenian legislation

Slovenian legislation is then imported into OntoGen…
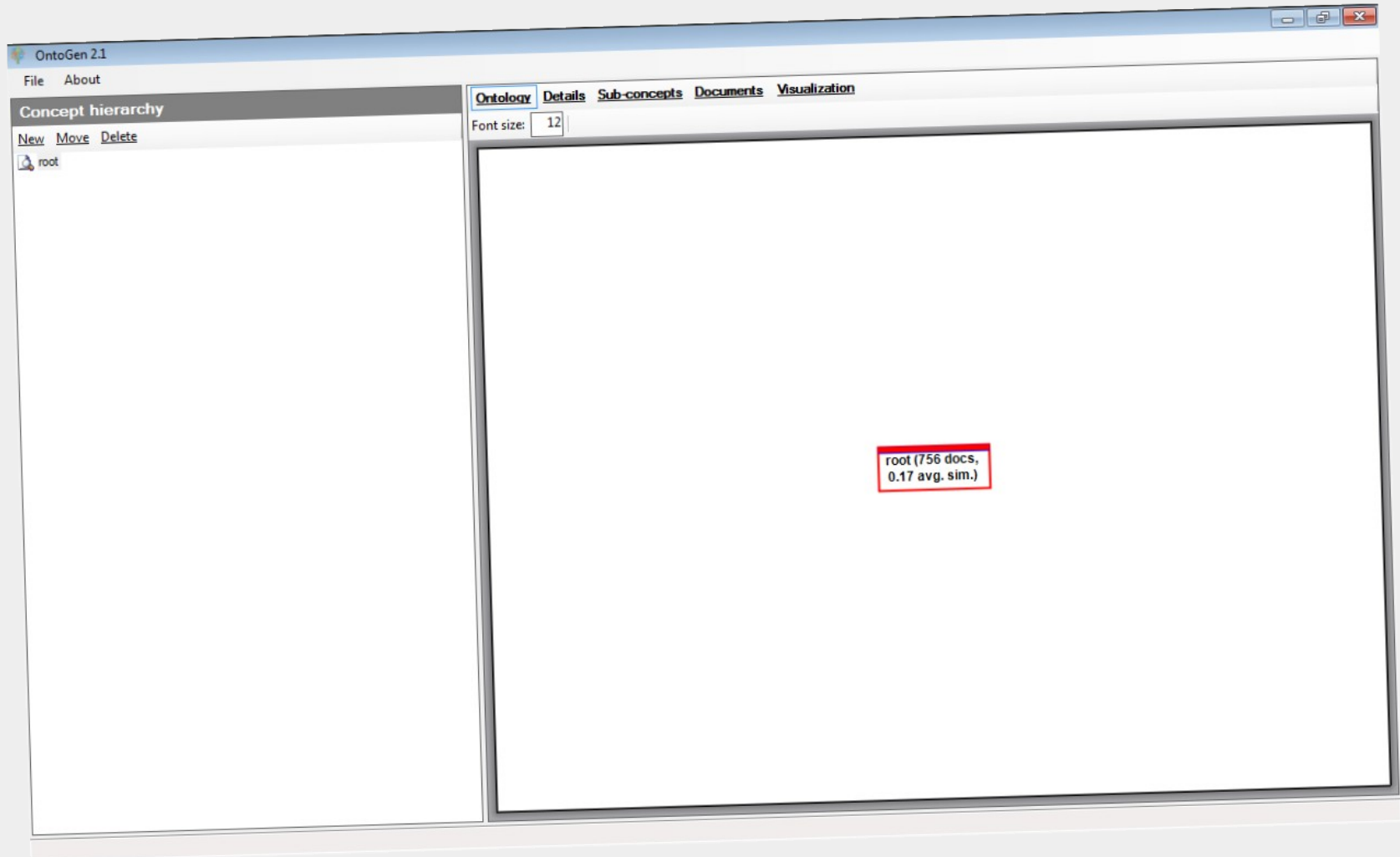
Text analysis of Slovenian legislation

OntoGen suggests us 8 subconcepts. We add subconcept named "tax"...

Text analysis of Slovenian legislation

For this subconcept we can preview documents included into this concept. We can also add or remove documents…

Text analysis of Slovenian legislation

Ontogen creates visualisation of the documents within the same concept...
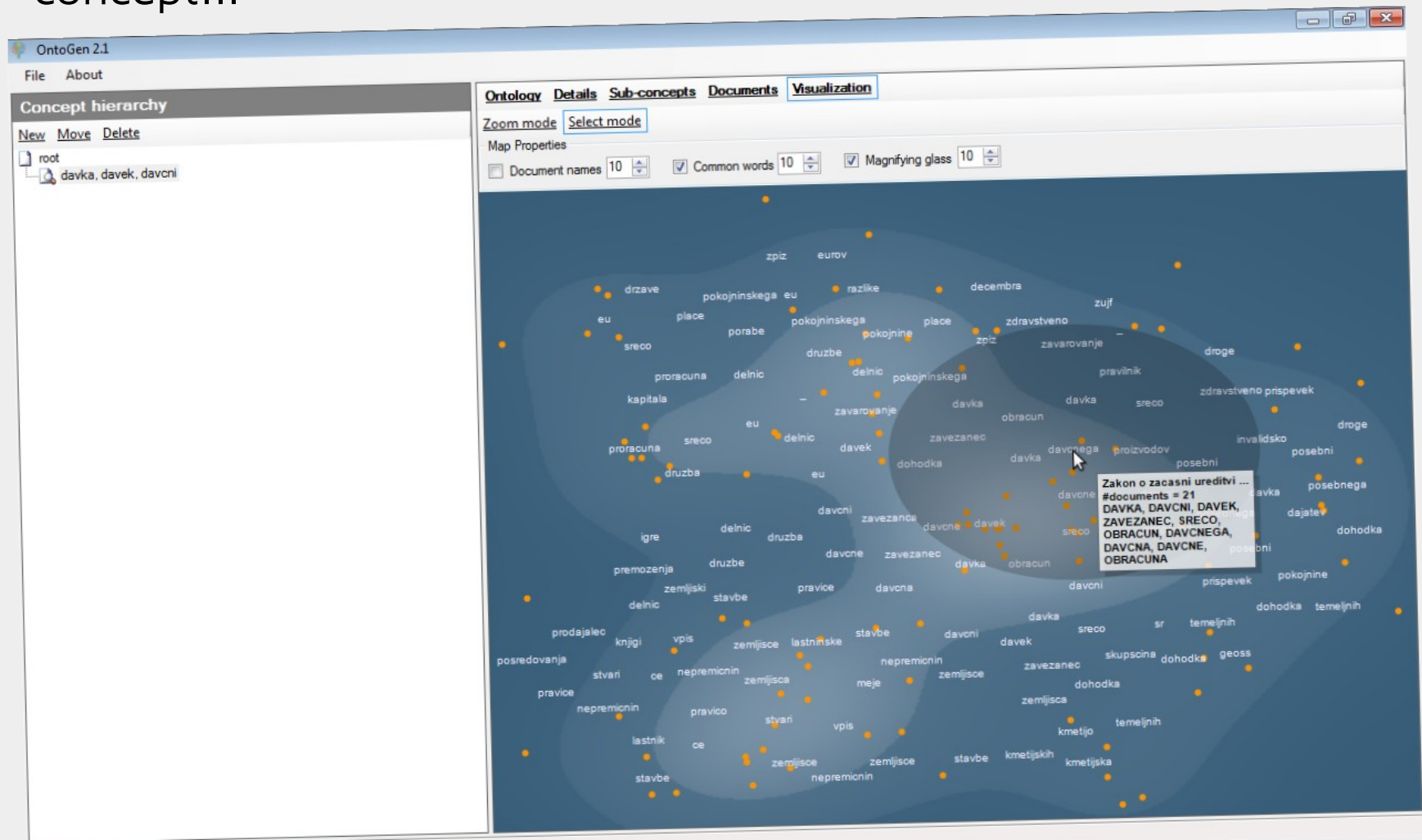
Text analysis of Slovenian legislation

We can view concept details...

Keywords, which separate the selected concept from other concepts most

Text analysis of Slovenian legislation

We can zoom in the concept...

Text analysis of Slovenian legislation

We can add new (sub)concept by learning...

Text analysis of Slovenian legislation

We can add new (sub)concept by learning…

Text analysis of Slovenian legislation

Documents from concept »tax« are automatically classified to subconcept »land«.

Example: Anonymisation of court decisions

Tacita is a tool helping anonymisation of court decisions. Based on a machine learning, the system has been taught which words to remove from the text. System has currently 91,8% recall (i.e. true positives).

Anonymisation of court decisions

Classification of words in two classes with the method of **support vectors**.
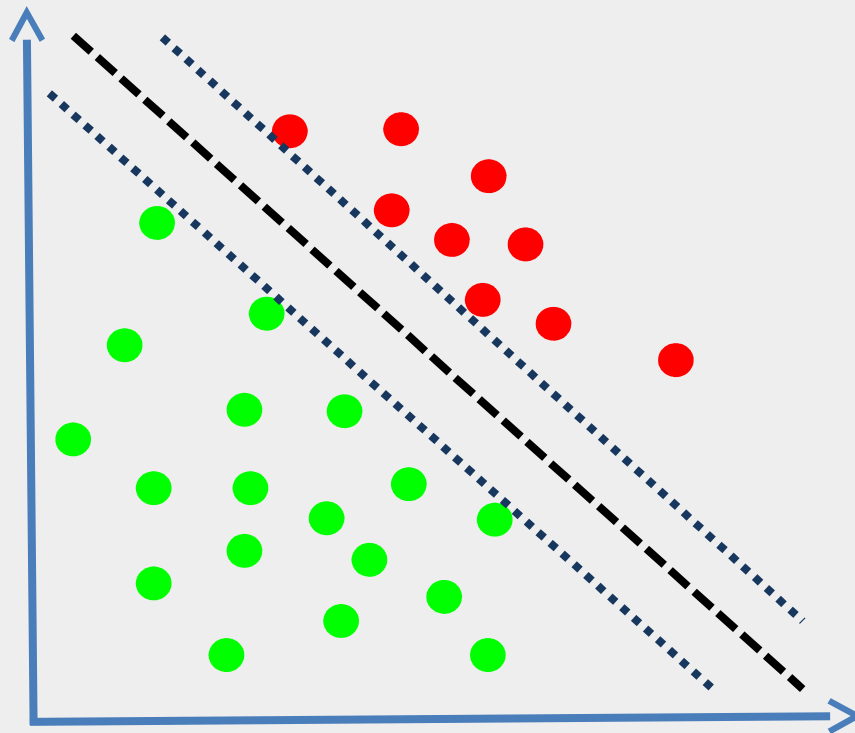
Linear distinguishable example

Word is represented as m-dimensional vector of several **features** (in our case m ~ 600).

Method tries to find an optimal **hyperplane**, which segregates points (words represented by vectors) in this (mapped) place.

Example: Geolocation of photographs

Geolocation of photographs

- Photographs taken from mobile phones have embed several metadata, including GPS location (usually in WGS 84 format).

- So we grab all data from the JPG images in a directory:
```
for file in *.jpg; do exiftool -c "%.6f" $file | grep "GPS
Position" | awk -v filename="$file" '{print filename, $4, $6}' | tr
[:blank:] \\t >> fotografije.csv; done;
```

- ... import it into a PostgreSQL database:
```
create table fotografije (filename text, latitude numeric,
longitude numeric);
\COPY fotografije from 'fotografije.csv' delimiter E'\t';
```

- ...and add geolocation (postGIS) info:
```
select AddGeometryColumn('fotografije', 'geolokacija', 4326,
'POINT', 2);
update fotografije set geolokacija =
ST_SetSrid(ST_GeomFromText('POINT(' || longitude || ' ' || latitude
|| ')'), 4326);
```

Geolocation of photographs
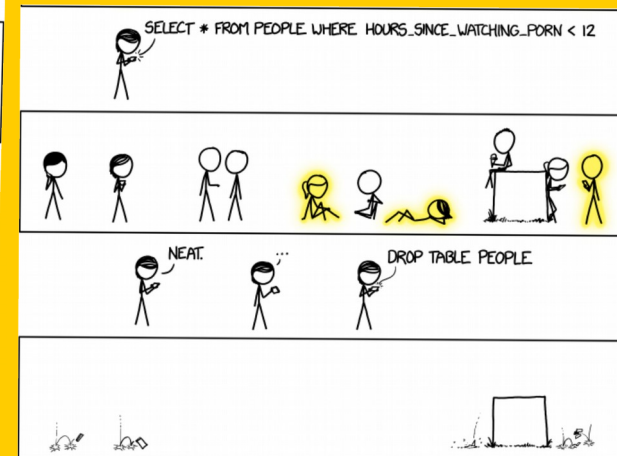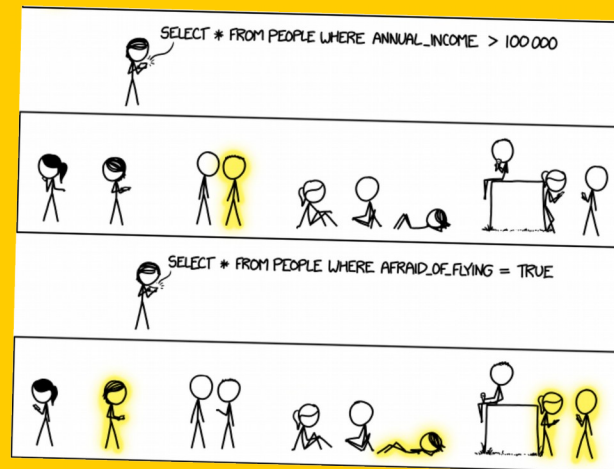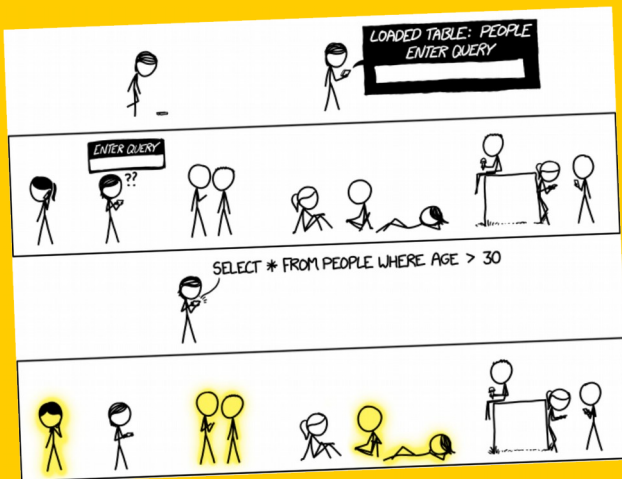
...and visualise data with opensource QGIS!
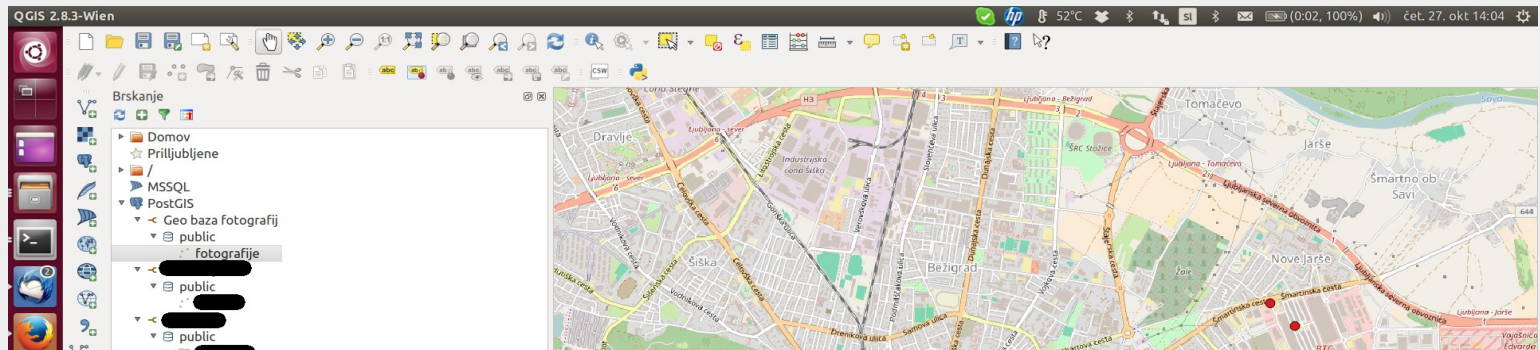
With PostGIS we can also do some nice locational analysis:

```
select ime_obcine, count(*) as
num_of_photos from fotografije, obcine
where (geolokacija_slo && geom) and
(ST_Contains(geom, geolokacija_slo))
group by ime_obcine;
```

```
ime_obcine | num_of_photos
-----------------+----------------
LJUBLJANA   |            86
KRANJ       |             5
```

SELECT *questions* FROM *audience*;

Cartoon: (CC) xkcd.com

Matej Kovačič
matej.kovacic@ijs.si

Centre for Knowledge Transfer in Information Technologies
http://ct3.ijs.si/

Artificial Intelligence Lab
http://ailab.ijs.si