Extracting Structured Data from Unstructured [Open] Data

Institute of Mathematics and Computer Science, University of Latvia

Uldis Bojārs – Renārs Liepiņš

SharePSI Krems workshop – 21-May-2015

Intro

- Goal: a Discussion
- Motivation:
  - countries with a smaller amount of "official" open data
  - content already published: text, webpages, ...
- Unstructured / Semi-structured content
  - may hold valuable [open] data

Unstructured – Semi-Structured – **Datasets**

Unstructured – **Semi-Structured** – Datasets

- Information already available on public sector organization sites

- BUT it is text, files or HTML only (!)
  - legislation, public procurement, MP votes, ...
  - this content has *some* structure

- Web scraping / Text processing
  - Open Data volunteers leading the way

Even data catalogues may require scraping:

OpenDataMonitor.eu study results – presented by Amanda Smith (@ayymanduh)
photo: https://twitter.com/Toon/status/601049078902980609

Balsošanas rezultāti

par 90, pret 0, atturas 0

Datums: **11/10/2011 10:03:27 AM**

Balsošanas motīvs: **Grozījumi Izglītības likumā (Saeimas kārtības ruļļa 39.p. kārtībā) (41/Lp11), 1.lasījums**

Grupēt: (•) pēc balsojuma rezultāta ( ) alfabētiskā kārtībā

	Vārds	Frakcija	Balss		Vārds	Frakcija	Balss
	PAR: 90			51.	Inese Lībiņa-Egnere	ZRP	Par
1.	Valērijs Agešins	SC	Par	52.	Ingmārs Līdaka	ZZS	Par
2.	Arvils Ašeradens	VIENOTĪBA	Par	53.	Igors Meļņikovs	SC	Par
3.	Dzintars Ābiķis	VIENOTĪBA	Par	54.	Sergejs Mirskis	SC	Par
4.	Solvita Āboltiņa	VIENOTĪBA	Par	55.	Ināra Mūrniece	VL-TB/LNNK	Par
5.	Aija Barča	ZZS	Par	56.	Romāns Naudiņš	VL-TB/LNNK	Par
6.	Andris Bērziņš	ZZS	Par	57.	Vladimirs Nikonovs	SC	Par
7.	Guntars Bilsēns	ZRP	Par	58.	Ņikita Ņikiforovs	SC	Par
8.	Inita Bišofa	ZRP	Par	59.	Klāvs Olšteins		Par
9.	Inga Bite	ZRP	Par	60.	Vitālijs Orlovs	SC	Par
10.	Raivis Blumfelds	VL-TB/LNNK	Par	61.	Jānis Ozoliņš	ZRP	Par
11.	Andris Buiķis	VIENOTĪBA	Par	62.	Imants Parādnieks	VL-TB/LNNK	Par
12.	Boriss Cilevičs	SC	Par	63.	Igors Pimenovs	SC	Par
13.	Einārs Cilinskis	VL-TB/LNNK	Par	64.	Vineta Poriņa	VL-TB/LNNK	Par
14.	Irina Cvetkova	SC	Par	65.	Sergejs Potapkins	SC	Par
15.	Ingmārs Čaklais	VIENOTĪBA	Par	66.	Dzintars Rasnačs	VL-TB/LNNK	Par
16.	Lolita Čigāne	VIENOTĪBA	Par	67.	Romualds Ražuks	ZRP	Par
17.	Edmunds Demiters	ZRP	Par	68.	Vladimirs Reskājs	SC	Par
18.	Sergejs Dolgopolovs	SC	Par	69.	Ivans Ribakovs	SC	Par
19.	Jānis Dombrava	VL-TB/LNNK	Par	70.	Dmitrijs Rodionovs	SC	Par
20.	Vjačeslavs Dombrovskis	ZRP	Par	71.	Artūrs Rubiks	SC	Par
21.	Raivis Dzintars	VL-TB/LNNK	Par	72.	Raimonds Rubiks	SC	Par

PAZIŅOJUMS PAR IEPIRKUMA PROCEDŪRAS REZULTĀTIEM

Publicēšanas datums: 11/05/2015

I IEDAĻA: PASŪTĪTĀJS

I.1. Nosaukums/vārds, adrese un kontaktpunkts (-i)

Pilns nosaukums, reģistrācijas numurs: Jelgavas novada pašvaldība, 90009118031

Pasta adrese: Pasta iela 37

Pilsēta/Novads: Jelgava | **Pasta indekss:** LV 3001 | **Valsts:** Latvija

Kontaktpunkts (-i): Jelgavas novada pašvaldība, Pasta iela 37,Jelgava
Kontaktpersonas vārds, uzvārds: Anželika Kanberga
Tālruņa numurs: 63012251

E-pasta adrese: anzelika.kanberga@jelgavasnovads.lv | **Faksa numurs:**

Interneta adreses
Vispārējā interneta adrese (URL): http://www.jelgavasnovads.lv
Pircēja profila adrese (URL): http://www.jelgavasnovads.lv

I.2. Pasūtītāja veids un galvenā (-ās) darbības joma (-as)

□ Ministrija vai jebkura cita valsts vai federāla iestāde, ieskaitot to reģionālās vai vietējās apakšnodaļas
□ Valsts vai federāla aģentūra/birojs
□ Reģionāla vai vietēja iestāde
□ Reģionāla vai vietēja aģentūra/birojs
□ Publisko tiesību subjekts
□ Eiropas institūcija/aģentūra vai starptautiska organizācija
■ Cits: *Pašvaldība*

■ Vispārēji sabiedriskie pakalpojumi
□ Aizsardzība
□ Sabiedriskā kārtība un drošība
□ Vide
□ Ekonomika un finanses
□ Veselība
□ Dzīvokļu un komunālā saimniecība
□ Sociālā aizsardzība

Public procurement – tender result publication
http://pvs.iub.gov.lv/show/404647

Issues

- Data Quality
  - multitude of quality issues in source documents
  - scraping may introduce errors
- Outdated information
  - if scraping done irregularly (e.g. at hackatons)
- Scraping should not be necessary
  - just publish as structured data

Best Practices

- Publish as Structured Data
  - instead of text, files, ...
- Pave the “Cow Paths”
  - prioritize datasets that people are scraping

Unstructured – Semi-Structured – Datasets

What can be done with Text

- Metadata Extraction
  - author; creation date; category
- Entity Extraction
  - people; organizations; place names
- Relationship Extraction
  - [Person A] works at [Organization 1]

Profile Extraction from News Articles

PERSONAS

Zatlers Valdis

Personas dati, ģimenes stāvoklis

dzimis 1955.gada 22.martā Rīgā

precējies, sieva Lilita Zatlere (dz.1953) - otrās laulības

bērni: 3

meita-Felicita Zatlere- Kotāne precējusies ar uzņēmēju Āri Kotānu

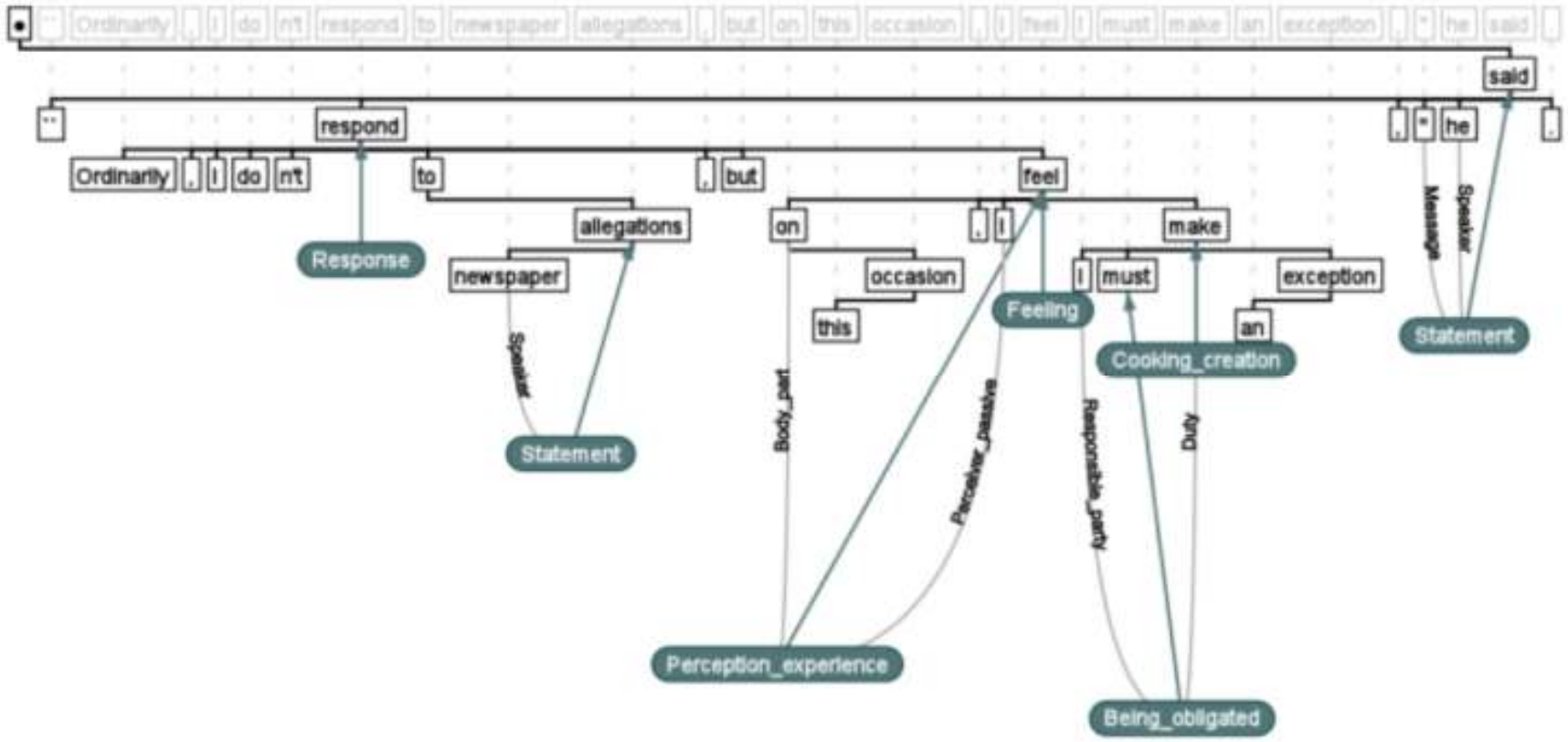
2008.28.jūlijā dzimst mazmeita

Izglītība

1973.- Rīgas 50.vidusskola

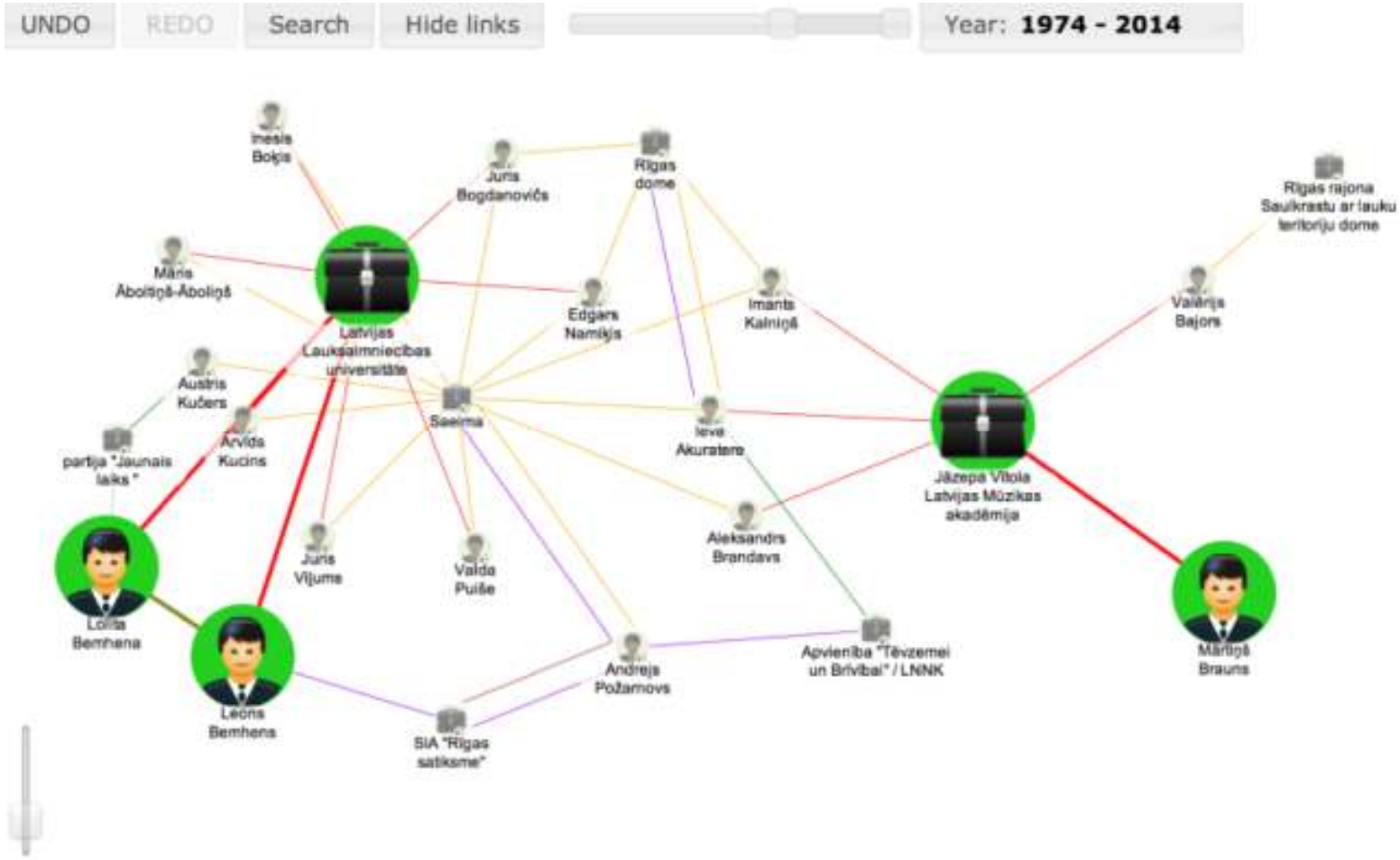
1979.- Rīgas Medicīnas institūts, ārsta specialitāte

1990.-1991.- izglītojies ASV (Yale University Visiting Scholar un Syracuse University, Visiting Orthopaedic Residency - 6 mēn.)

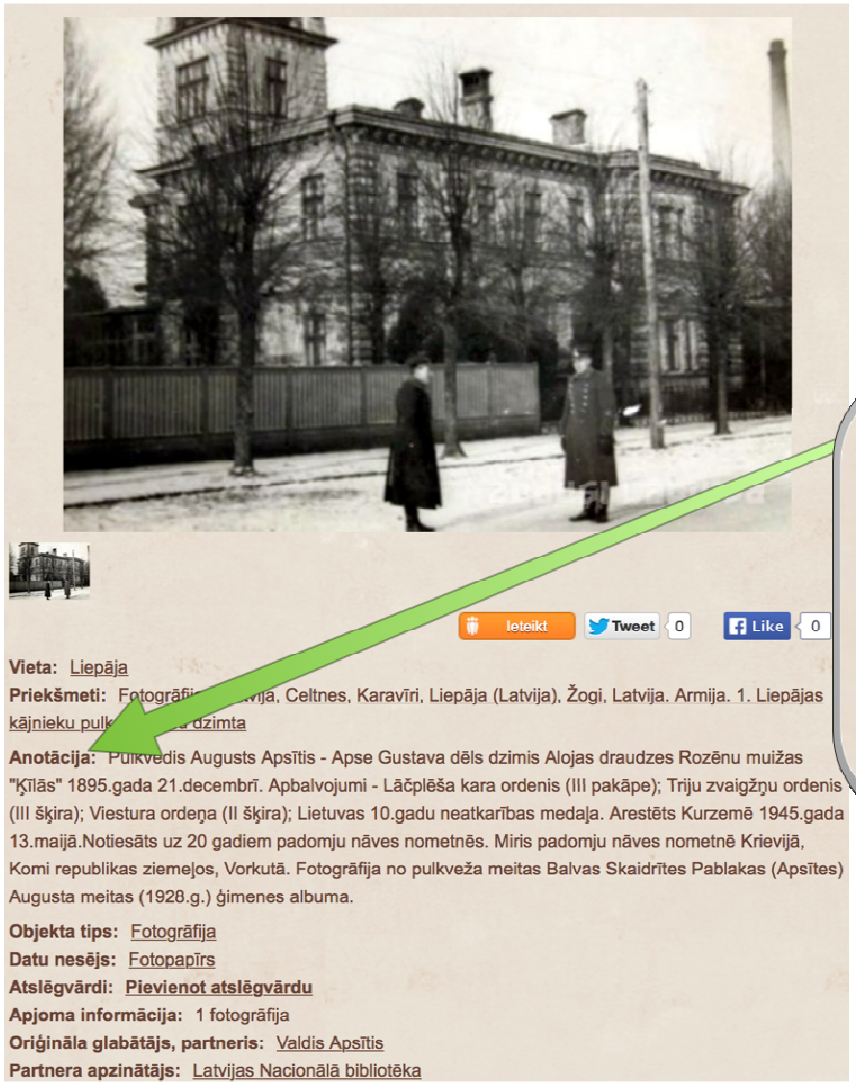


FrameNet CNL: A knowledge representation and information extraction language.
Barzdins, G. (2014)
http://c60.ailab.lv

Visualization of Person / Organization relationships



Cultural Heritage Data



Ieteikt Tweet 0 Like 0

Vieta: Liepāja

Priekšmeti: Fotogrāfi… …vija, Celtnes, Karavīri, Liepāja (Latvija), Žogi, Latvija. Armija. 1. Liepājas kājnieku pulk… …dzimta

Anotācija: Pulkvedis Augusts Apsītis - Apse Gustava dēls dzimis Alojas draudzes Rozēnu muižas "Ķīlās" 1895.gada 21.decembrī. Apbalvojumi - Lāčplēša kara ordenis (III pakāpe); Triju zvaigžņu ordenis (III šķira); Viestura ordeņa (II šķira); Lietuvas 10.gadu neatkarības medaļa. Arestēts Kurzemē 1945.gada 13.maijā.Notiesāts uz 20 gadiem padomju nāves nometnēs. Miris padomju nāves nometnē Krievijā, Komi republikas ziemeļos, Vorkutā. Fotogrāfija no pulkveža meitas Balvas Skaidrītes Pablakas (Apsītes) Augusta meitas (1928.g.) ģimenes albuma.

Objekta tips: Fotogrāfija

Datu nesējs: Fotopapīrs

Atslēgvārdi: Pievienot atslēgvārdu

Apjoma informācija: 1 fotogrāfija

Oriģināla glabātājs, partneris: Valdis Apsītis

Partnera apzinātājs: Latvijas Nacionālā bibliotēka

Anotācija: Pulkvedis Augusts Apsītis - A
"Ķīlās" 1895.gada 21.decembrī. Apbalvoj
(III šķira); Viestura ordeņa (II šķira); Lietu
13.maijā.Notiesāts uz 20 gadiem padomj
Komi republikas ziemeļos, Vorkutā. Foto

http://zudusilatvija.lv/objects/object/33199/

Discussion

- Anyone else doing this ?
- Should it be part of PSI ?
- Should we discuss this more ?