NACIONALNA I SVEUČILIŠNA KNJIŽNICA U ZAGREBU

Označivanje kao temelj jezika, učenja i obrade teksta

Marko Orešković

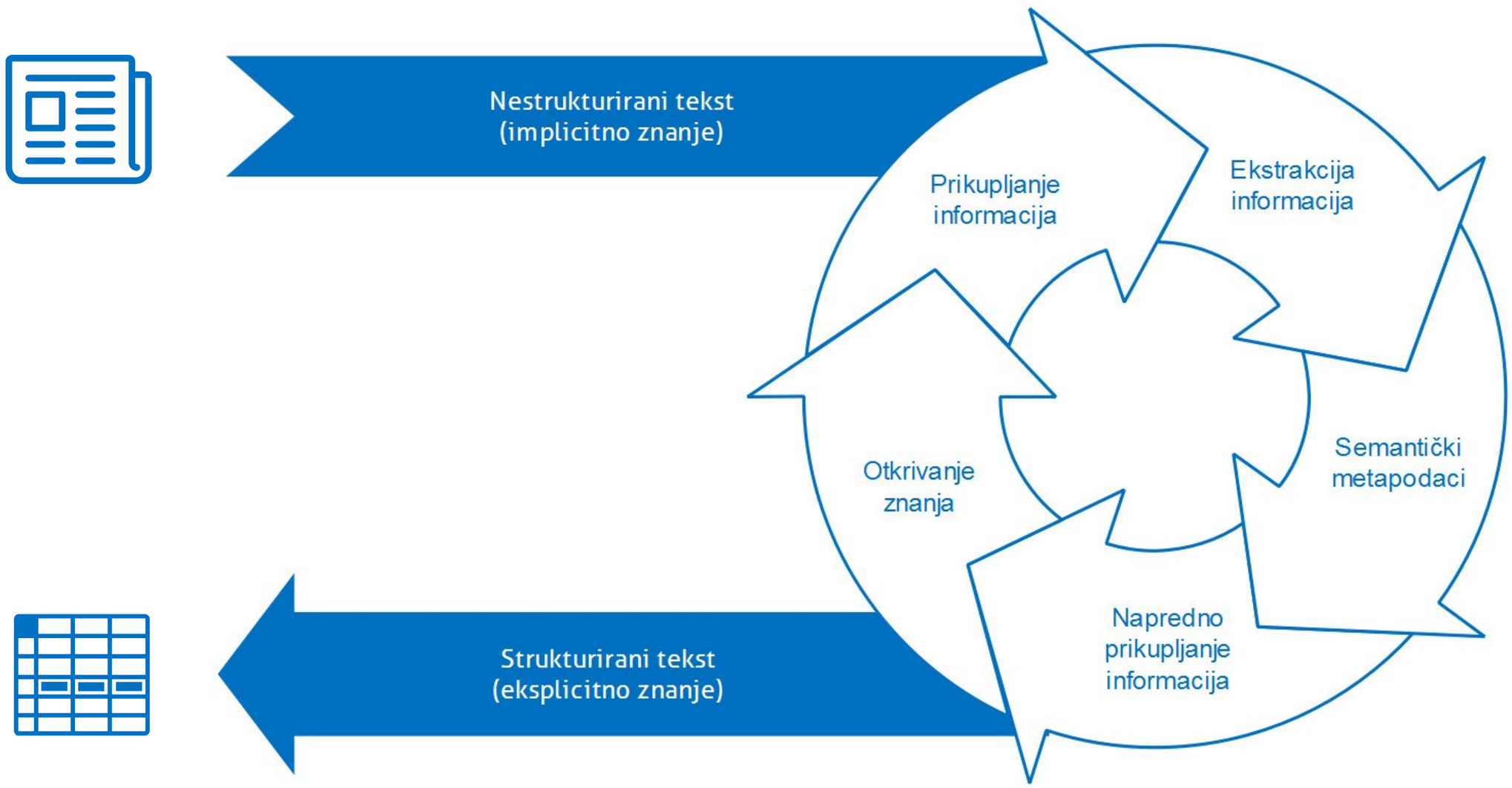
moreskovic@nsk.hr

Zagreb, 10. svibnja 2019.

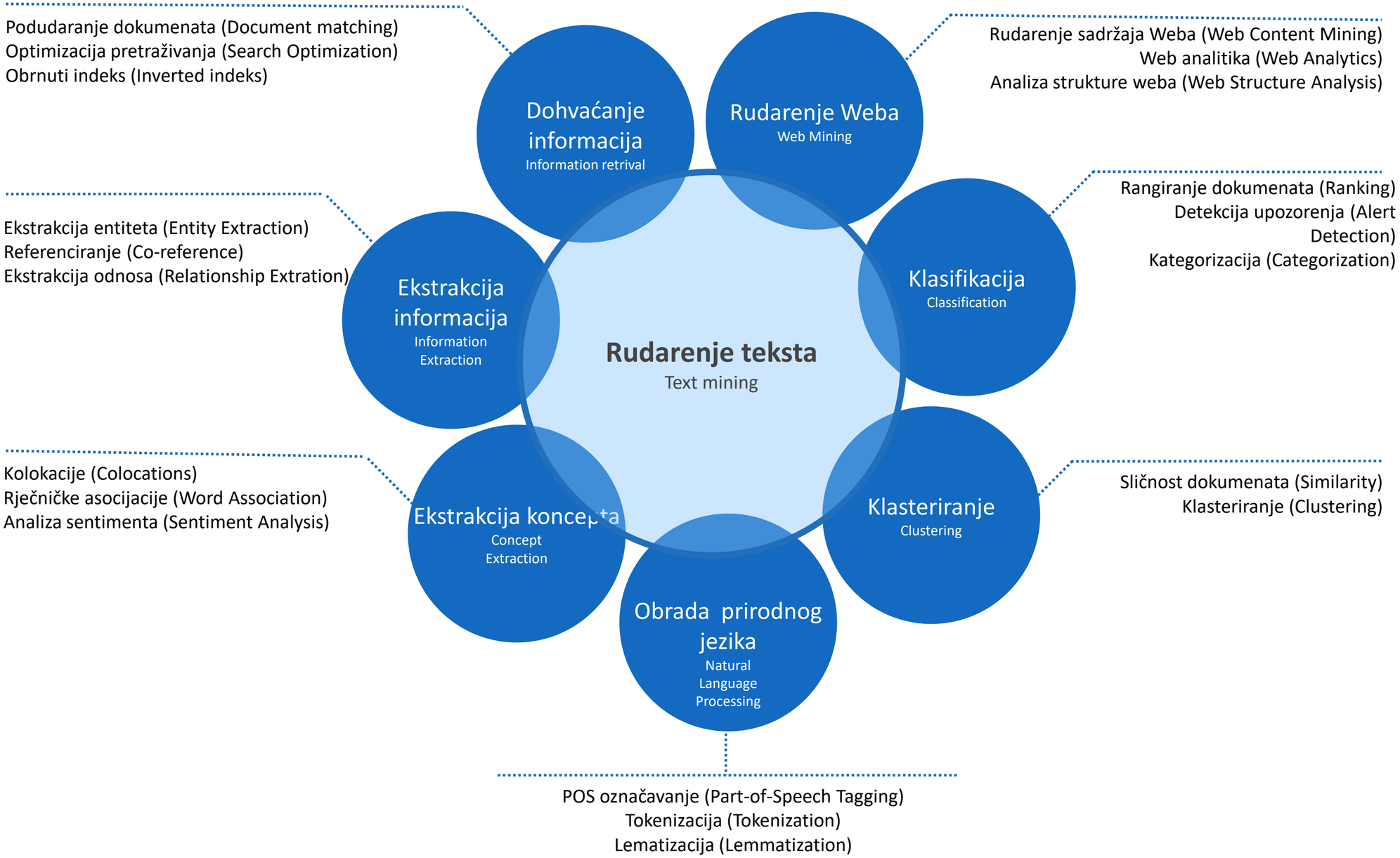
Sadržaj

Uvod

Kako dohvatiti (naj)više semantičke informacije iz digitaliziranog teksta ?

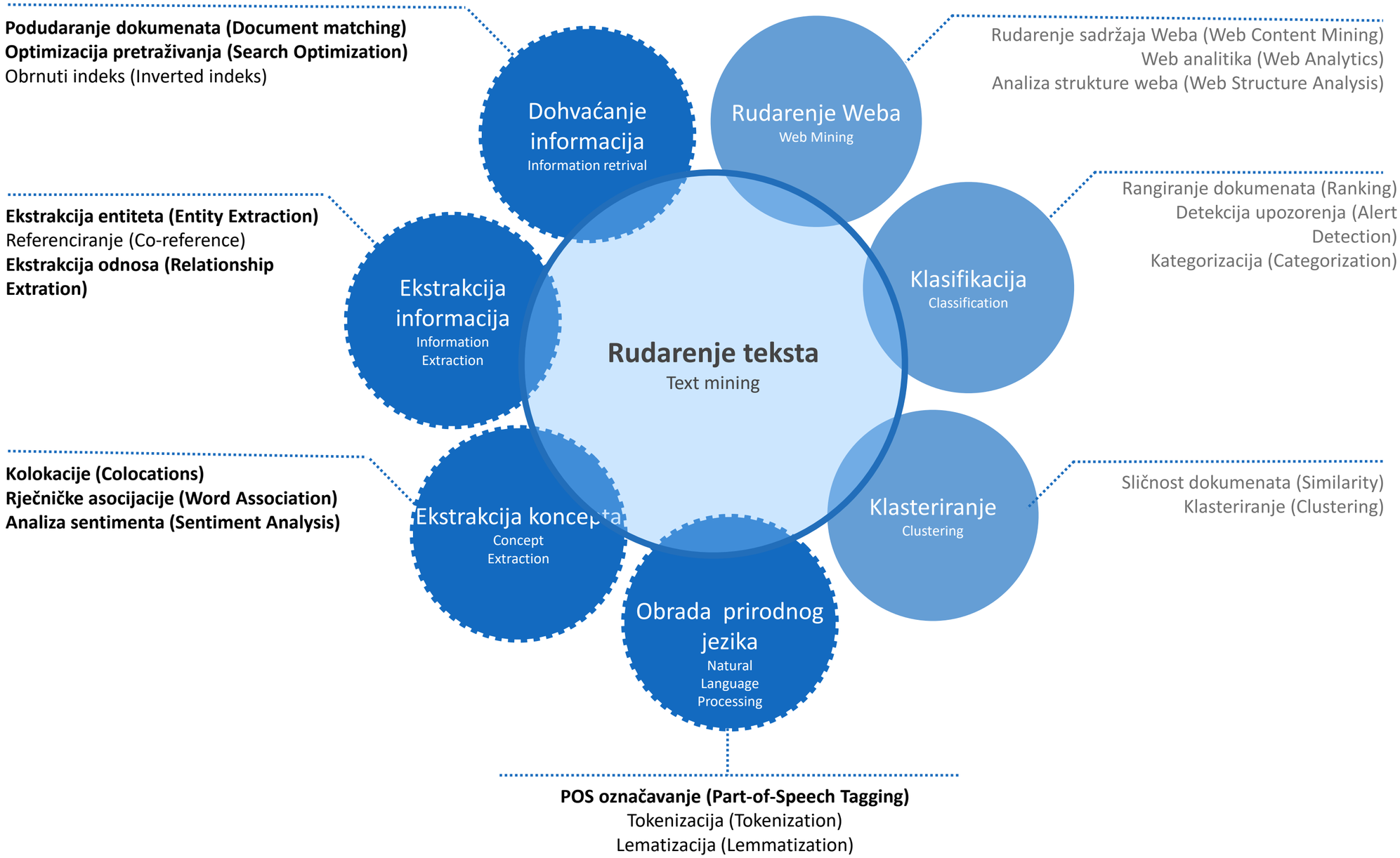


Rudarenje teksta – obrada prirodnog jezika



Stohastički vs. deterministički model

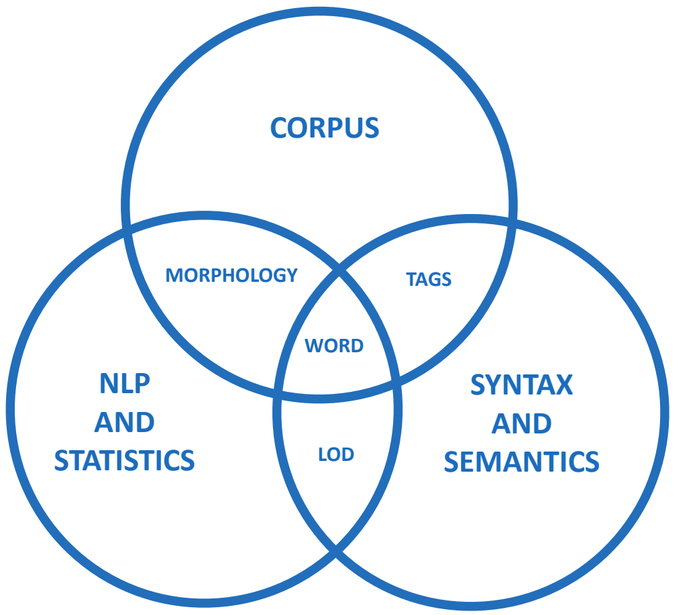
Rudarenje teksta – obrada prirodnog jezika

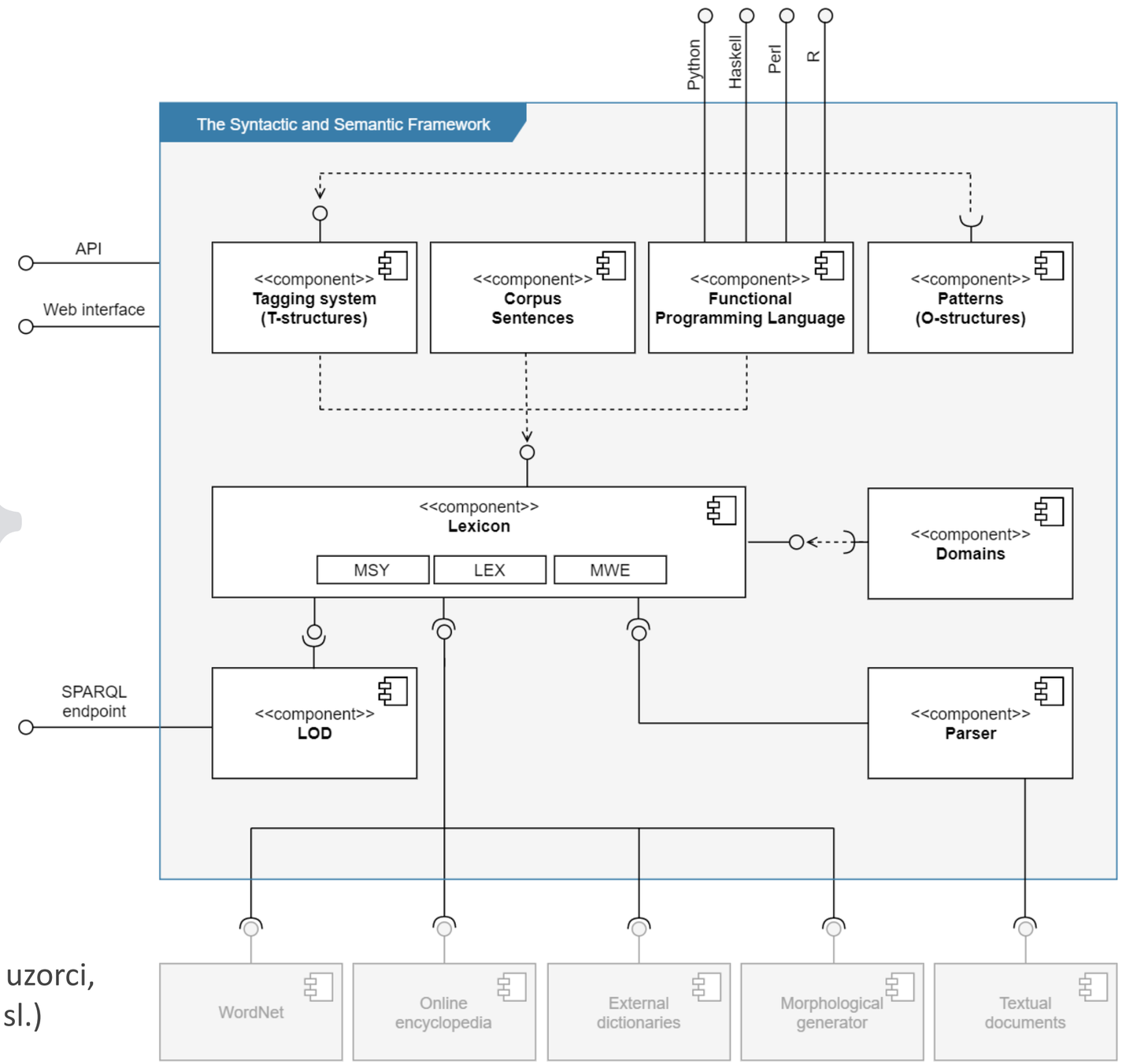


Stohastički vs. deterministički model

Deterministički računalni model prirodnog jezika

- U središtu modela je riječ
- Za strojnu obradu nužna su digitalna obilježja (tagovi)

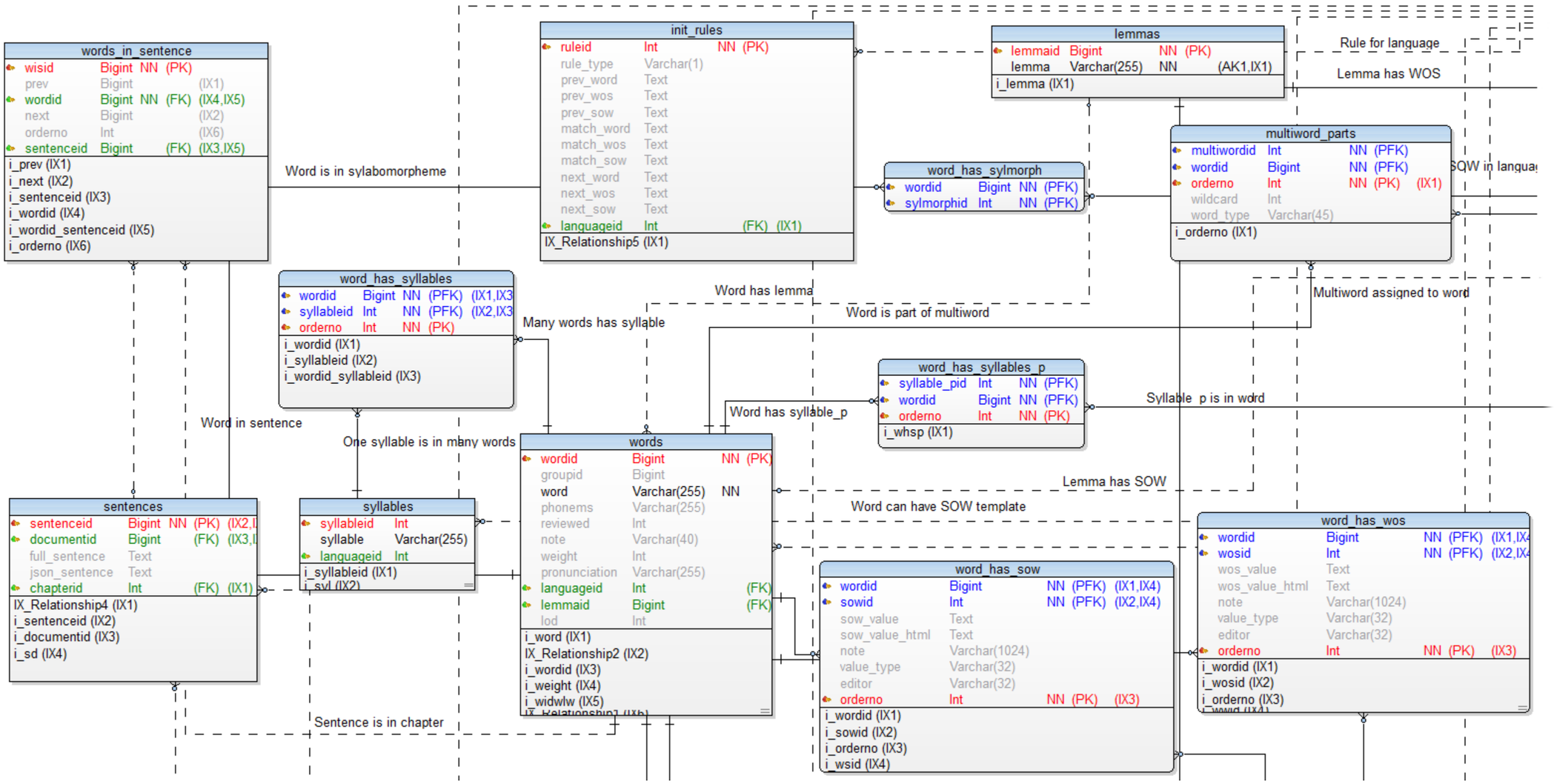




- Svaka razina ima svoja obilježja (npr. sintaksa: uzorci, funkcije /S-P-O/; semantika: sentiment, NER i sl.)

Realizacija modela

Konceptualni model pretvoren u relacijski model



Implementiran u MariaDB

Sadrži 40 tablica, 250 atributa, preko 200 indeksa (~ 5Gb podatkovnog prostora)

Javno dostupan

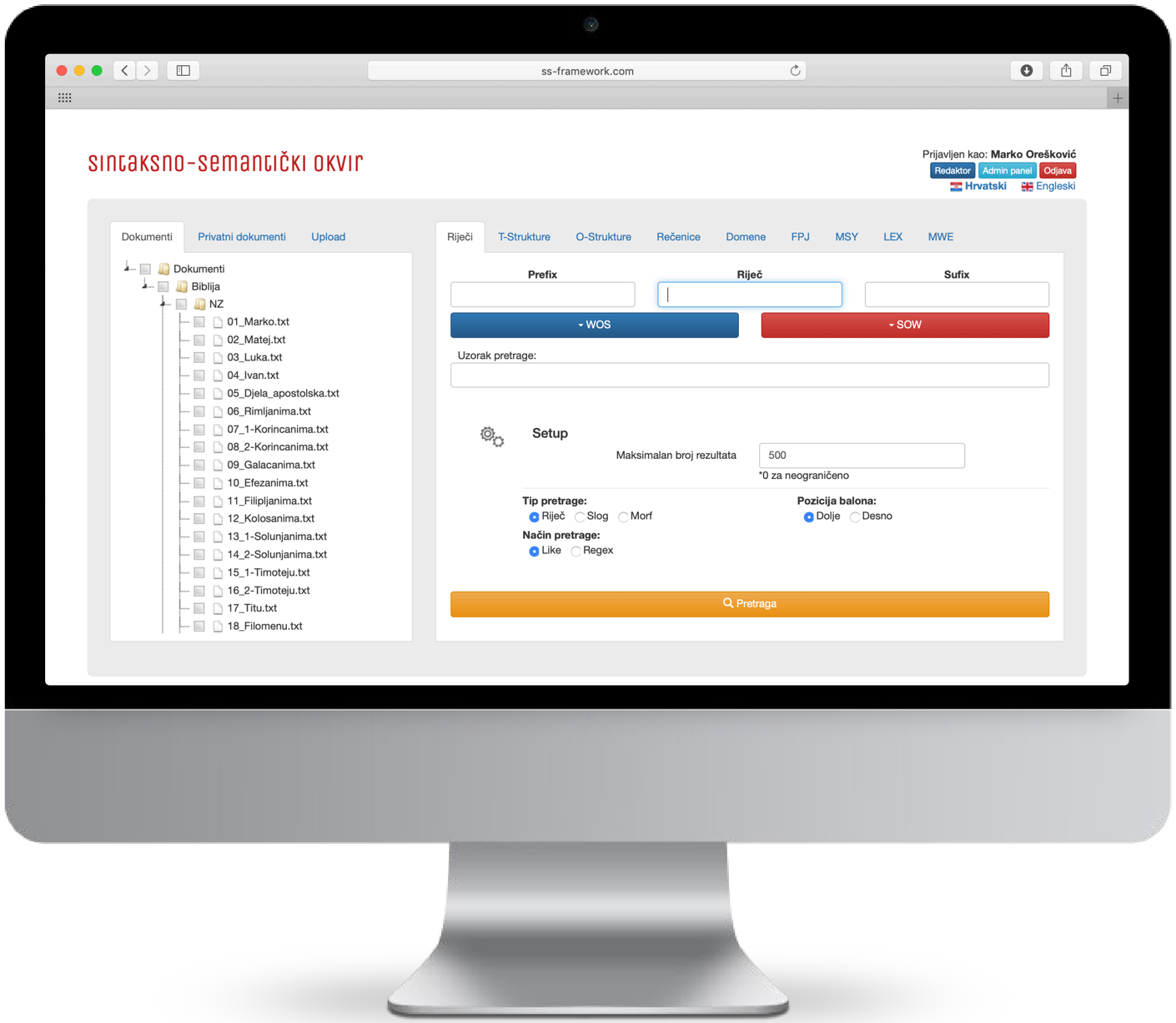
- Web aplikacija
- Javno dostupna:

  http://ssf.mathos.hr
- Frontend:

  Bootstrap, jQuery
- Backend:

  PHP, Python, MariaDB,

  Virtuoso triplestore



Stablo obilježja slično ontologijama

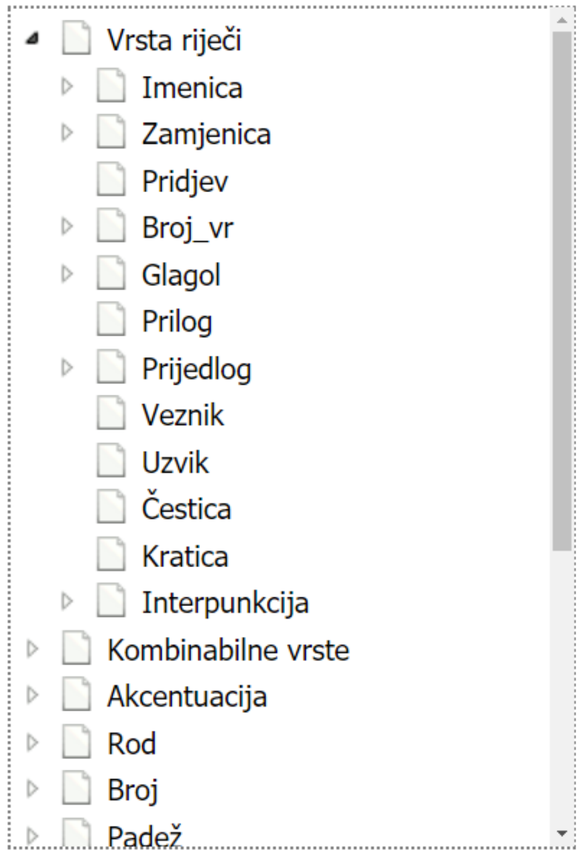
MULTEXT EAST

P	Atribut	Vrijednost	Kôd
0	KATEGORIJA	Imenica	N
1	Tip	opća	c
		vlastita	p
2	Rod	muški	m
		ženski	f
		srednji	n
3	Broj	jednina	s
		množina	p
4	Padež	nominativ	n
		genitiv	g
		dativ	d
		akuzativ	a
		vokativ	v
		lokativ	l
		instrumental	i
7	Živo	ne	n
		da	y

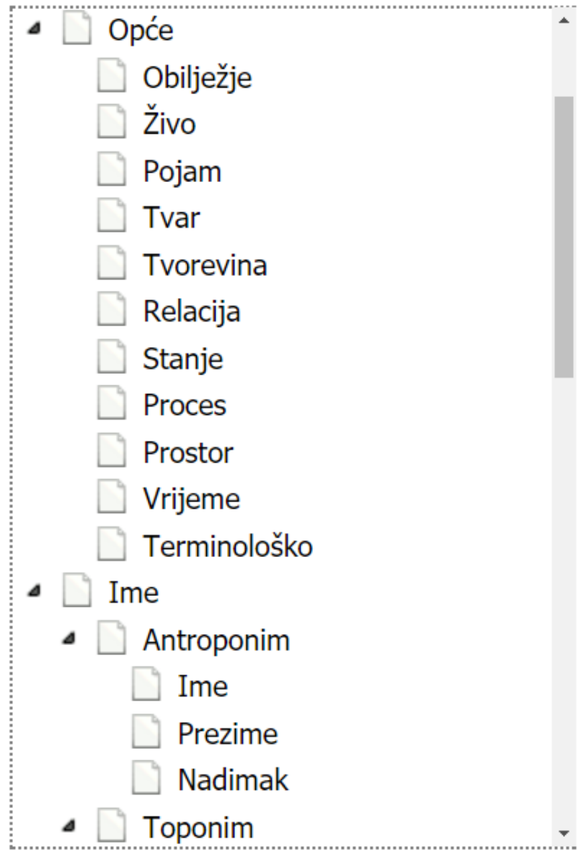
VS

T-STRUKTURE

WOS

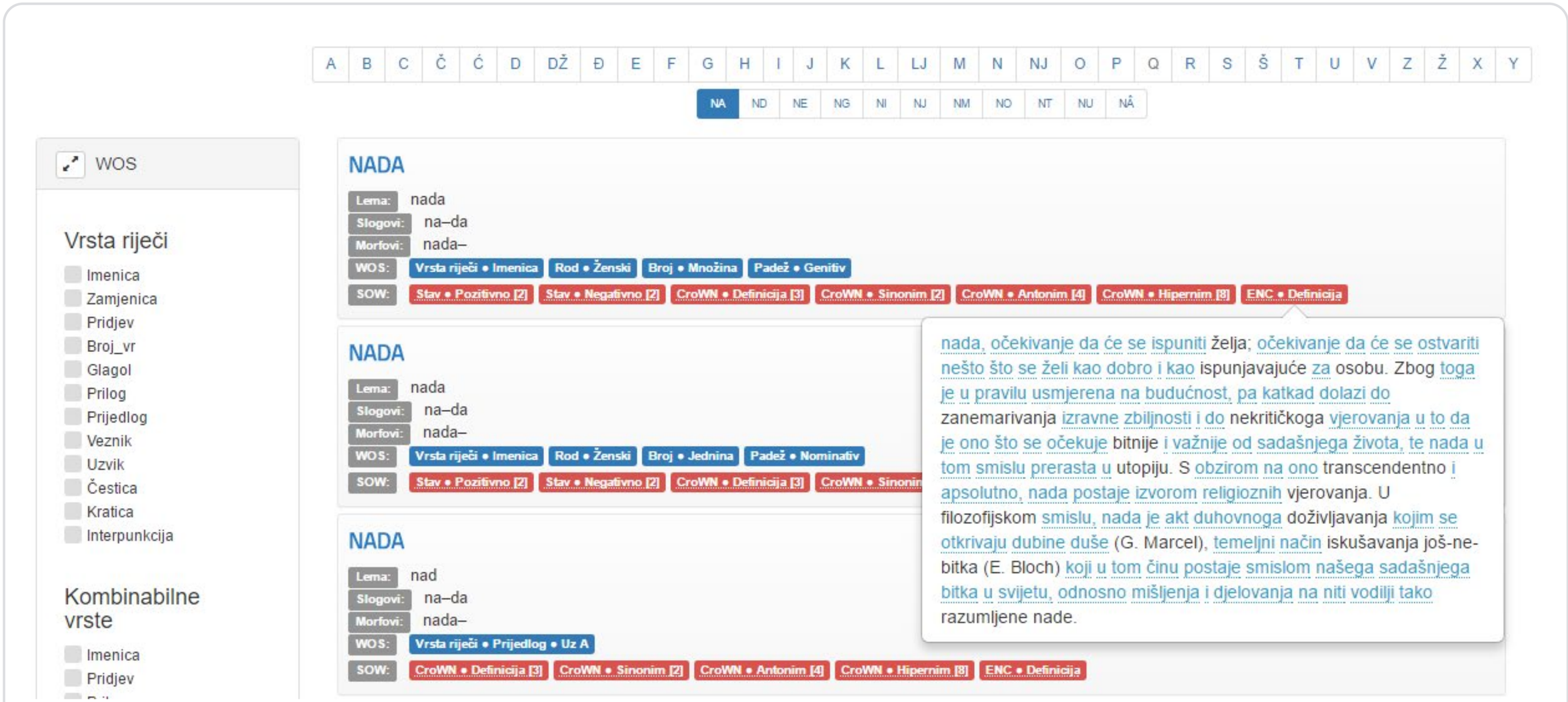


SOW



- WOS – word of speech (gramatička obilježja)
- SOW – semantic of word (semantička obilježja)

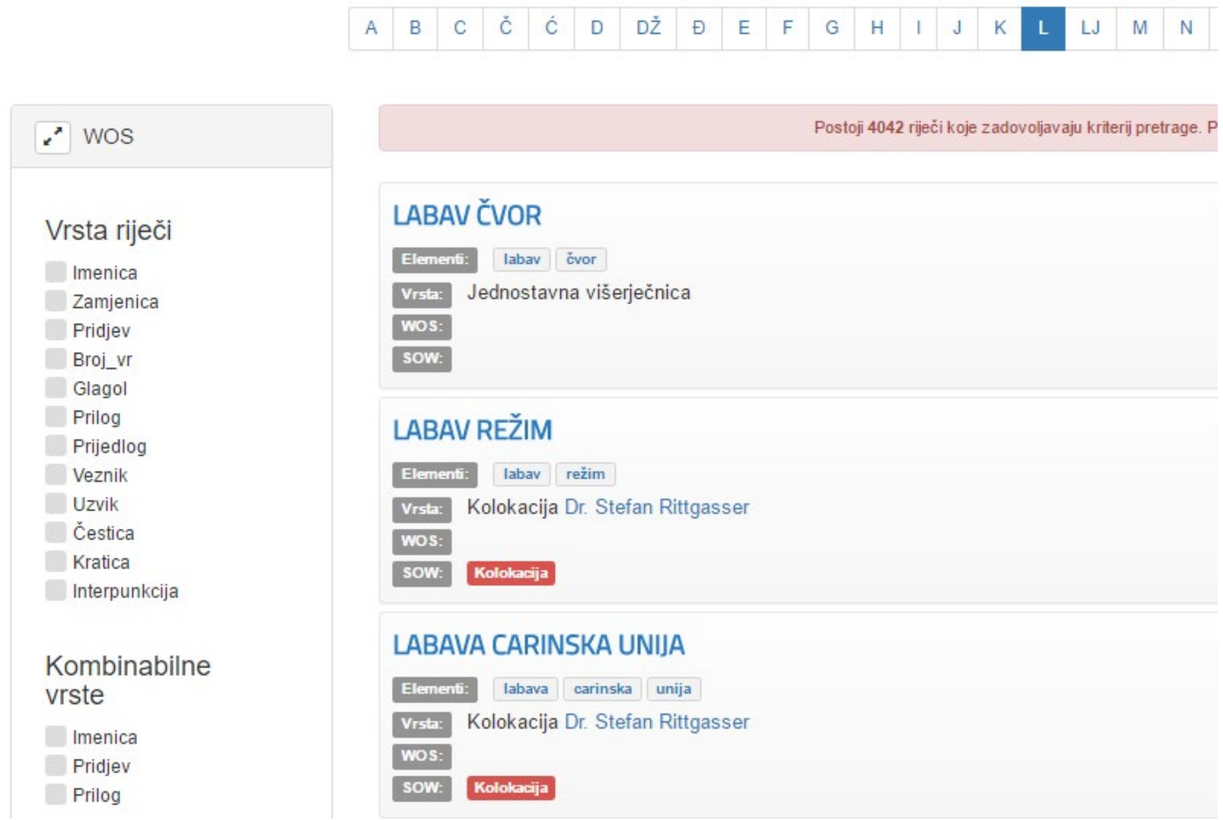
Povezanost riječi s repozitorijima i enciklopedijom



- HJP, LZMK, CroWN, Rječnik sinonima..
- Riječi iz definicije uz natuknice povezane u semantičku mrežu

Od atomarnih elemenata riječi do složenih izraza

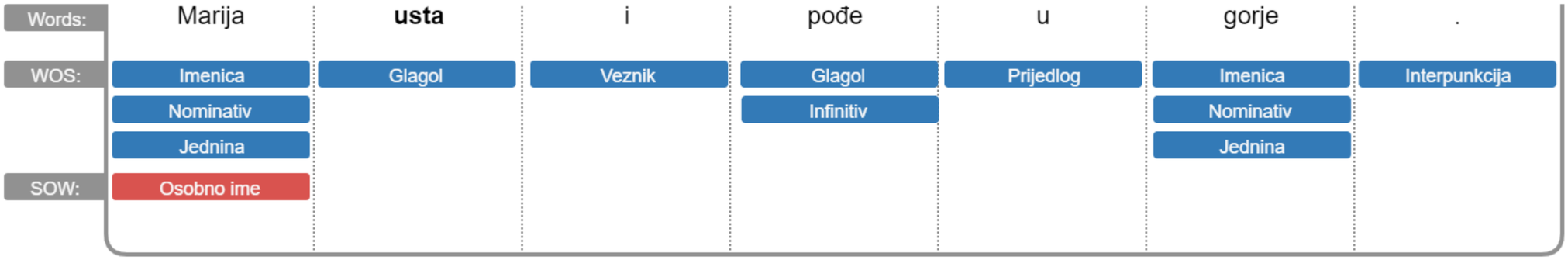
- Morfovi (2.118), slogovi (7.787), morfemi (796.448), višerječnički izrazi (121.771)

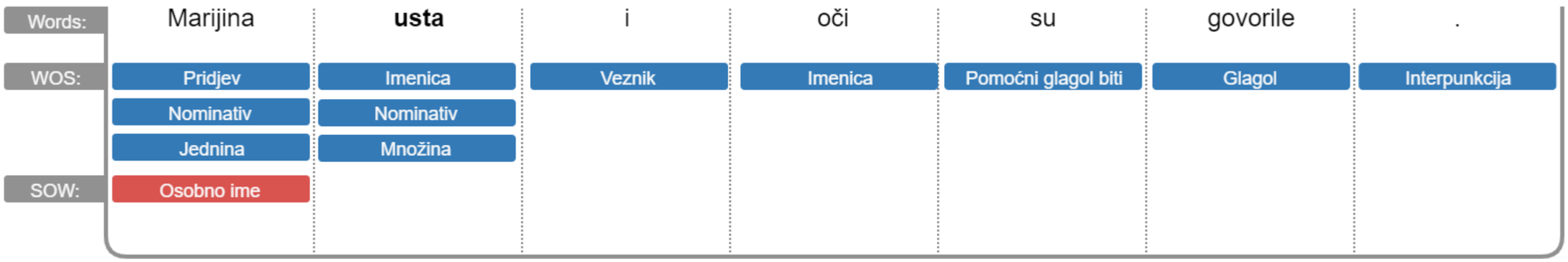


- Napredan način pretrage i filtriranja
- MSY: slogovi, morfovi, silabomorfemi
- MWE: kolokacije, frazemi, višerječnice

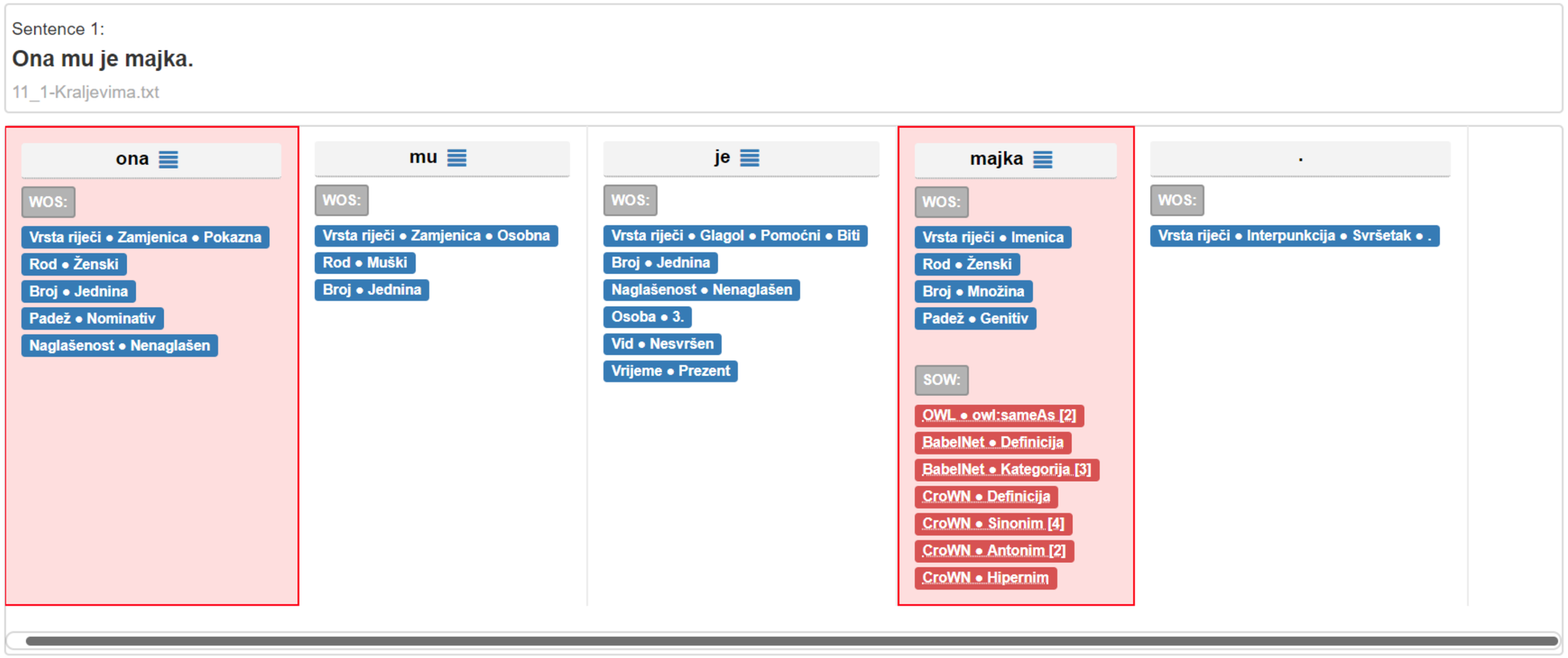
Sintaksni uzorci

Riješeni kompleksni problemi višeznačnosti





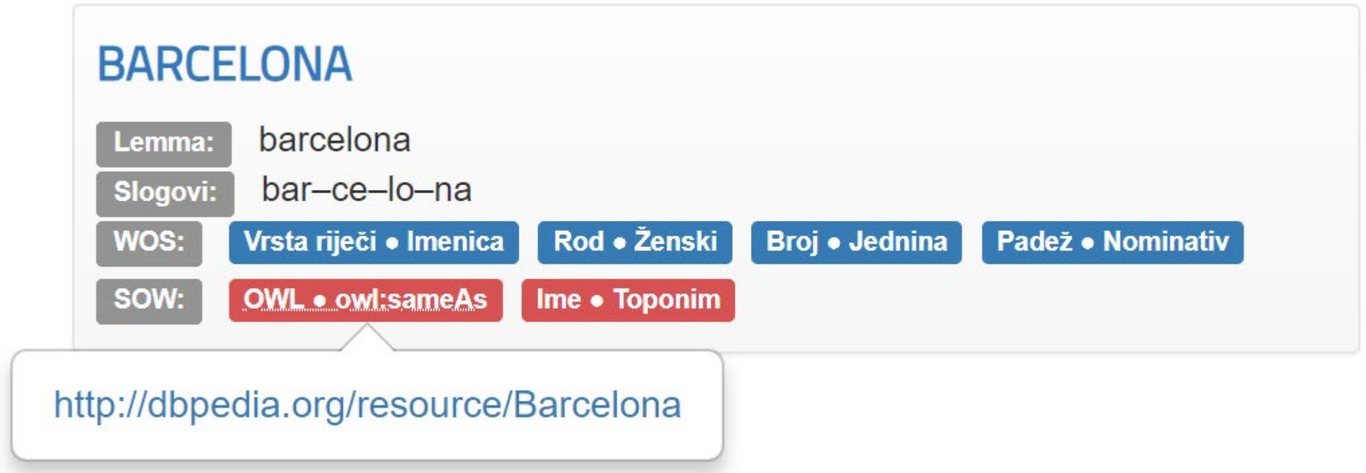
Izvlačenje sintaktičko-semantičke informacije iz teksta



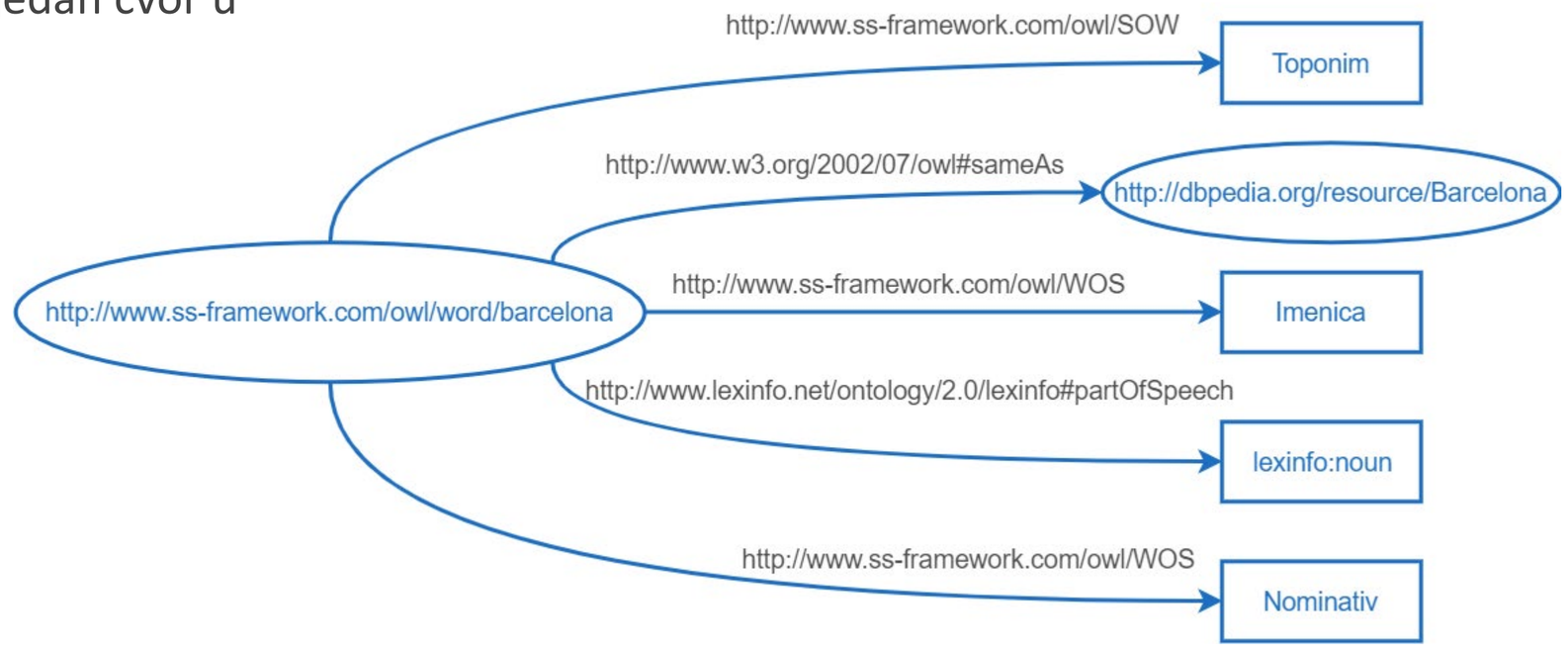
- Uz zadana WOS obilježja
- i/ili SOW obilježja
- i k tomu različitih tipova

Označivanjem do umrežene informacije

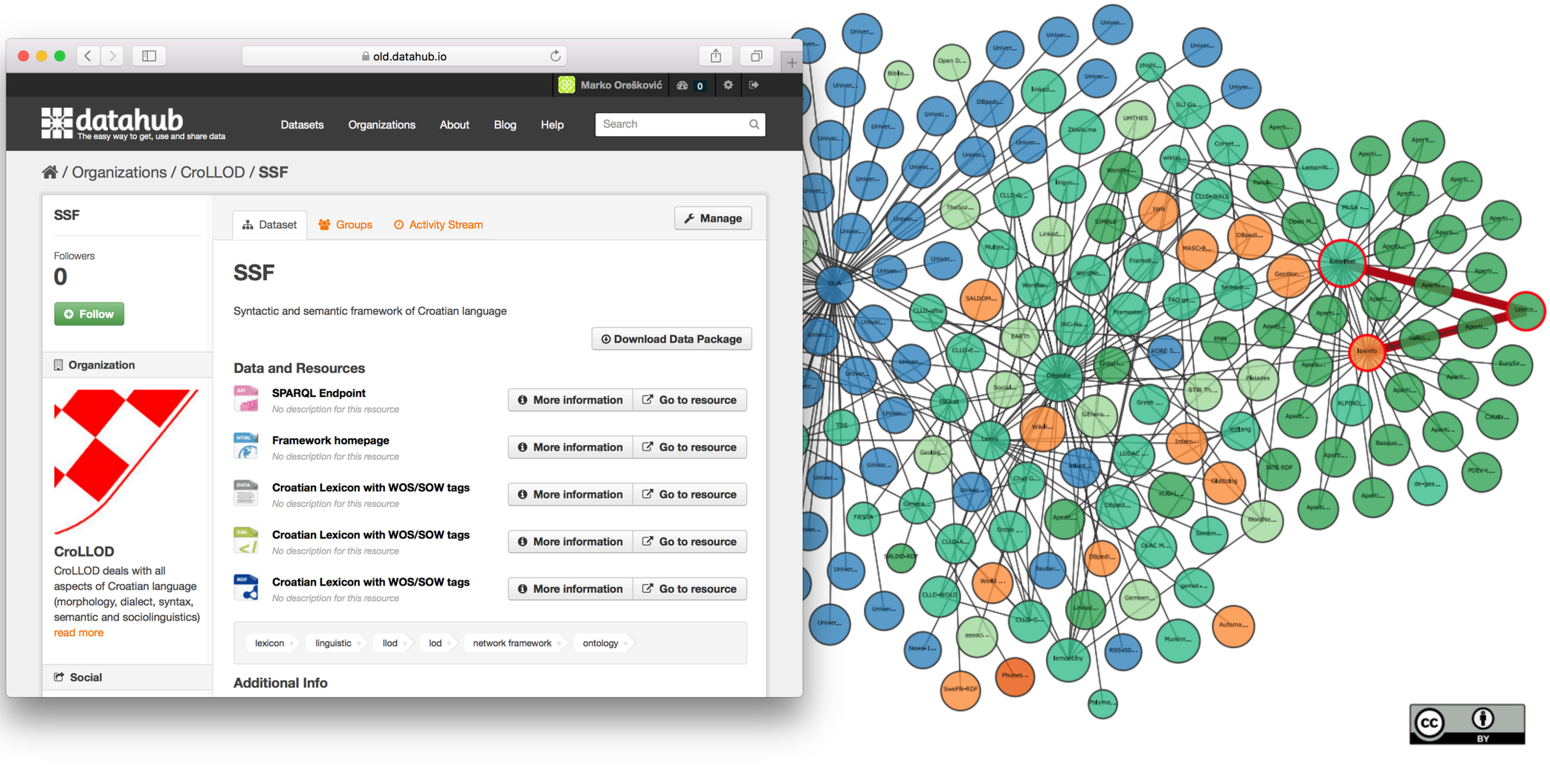
- Poseban tip SOW tagova **owl:sameAS** za povezivanje s vanjskim resursima
- Drugi WOS/SOW tagovi transformiraju se u RDF trojke



- Svaka riječ u SSF-u je jedan čvor u ontologiji



CroLLOD u svjetskoj globalnoj mreži



- Od travnja 2018, SSF leksikon je dio globalnog LOD oblakasa 70,366 trojaca, od kojih 67,717 je vezano na LexInfo, 35,687 na Princeton WordNet i 20,456 na BabelNet

Application programming interface (API)

- Integracija s vanjskim resursima
- REST API preko HTTP, primjer Python koda:

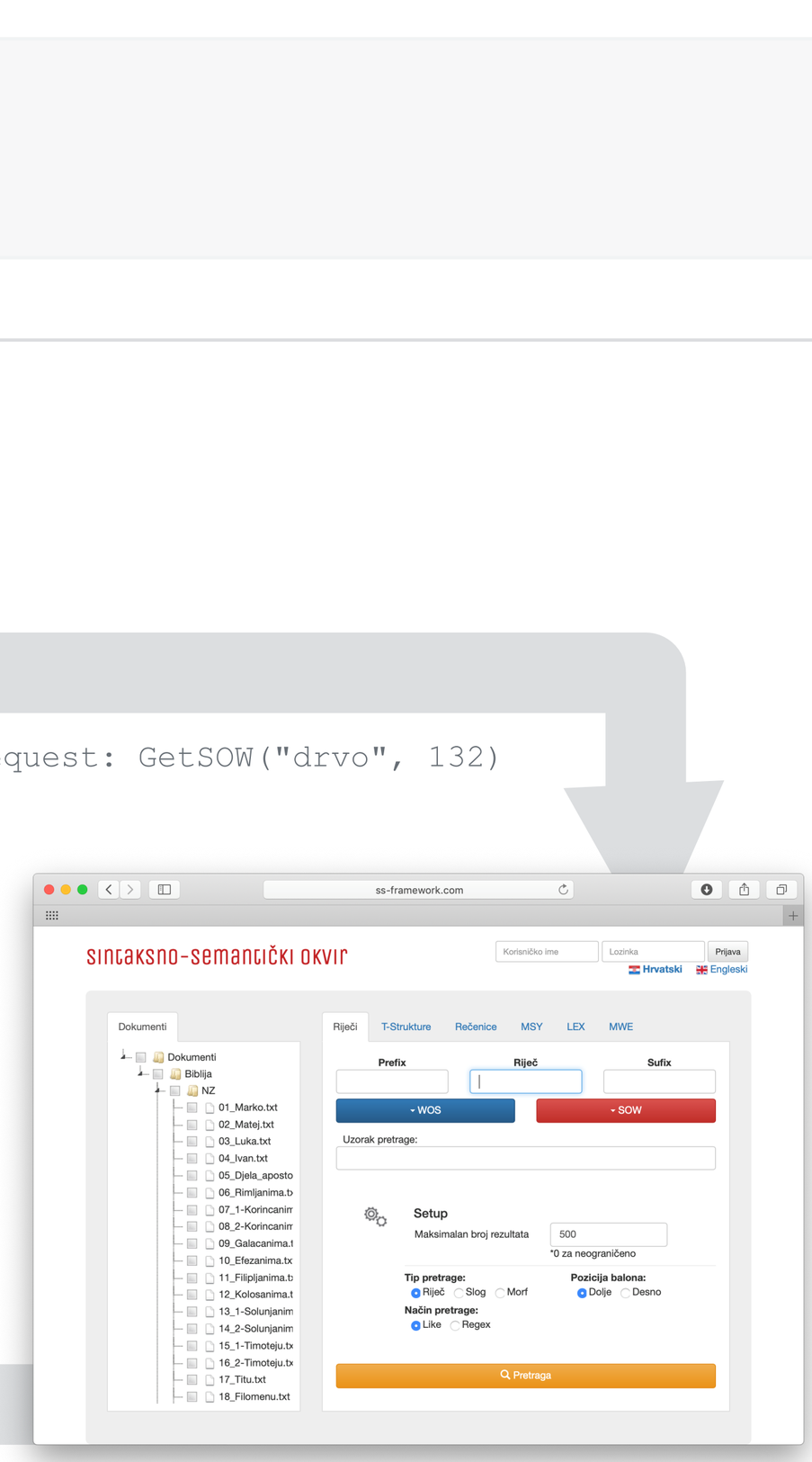
```python
# -*- coding: utf-8 -*-
import requests

# API key
key = "429ae4bfe8088f071abef86ac021653b"

# Execute in Python, Haskel, SPARQL, R
program = "Python"

# Code to be executed
code = "=ChangeTense('vidim plavu kuću', 'aorist')"
url = "http://www.ss-framework.com/api/fpj"

req = requests.post(url, data = {'apiKey':key, 'program':program, 'code':code})

# Set the output encoding to UTF-8
req.encoding = 'UTF-8'

# Print the output
print req.text
```

- Odgovor: `{"status":200, "result":"vidjeh plavu kuću"}`

Application programming interface (API)

Primjer: http://www.suncenaprozorcicu.com

HTTP Request: GetSOW("drvo", 132)



{"status":200, result: "http://www.ss-framework.com/images/drvo.png"}

HVALA

Marko Orešković
Varaždin, 02.06.2017.