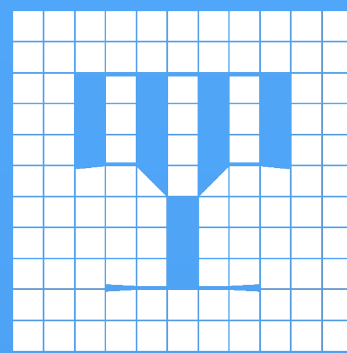




NACIONALNA I
SVEUČILIŠNA
KNJIŽNICA
U ZAGREBU

Označivanje kao temelj jezika, učenja i obrade teksta

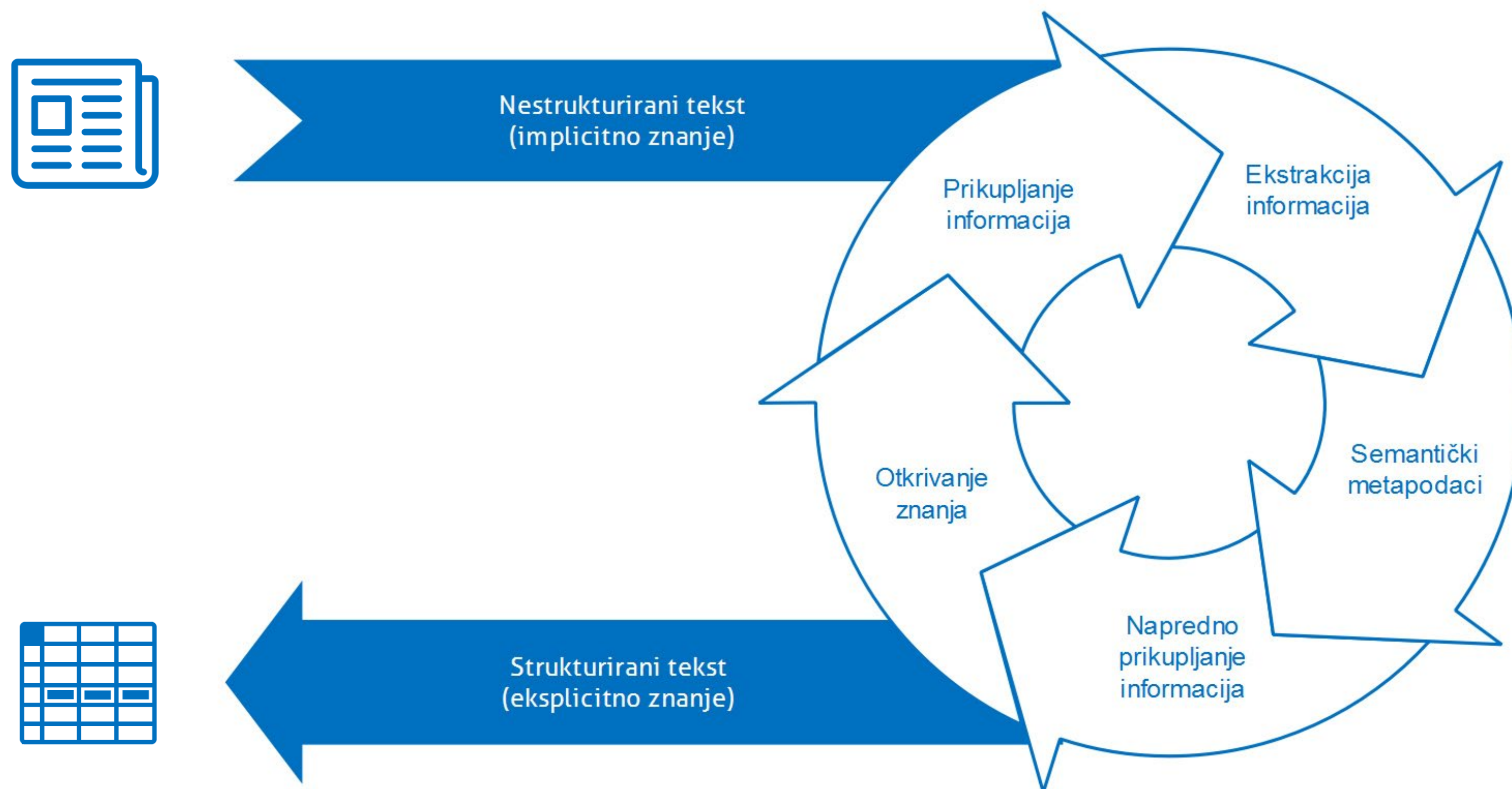
Marko Orešković

moreskovic@nsk.hr

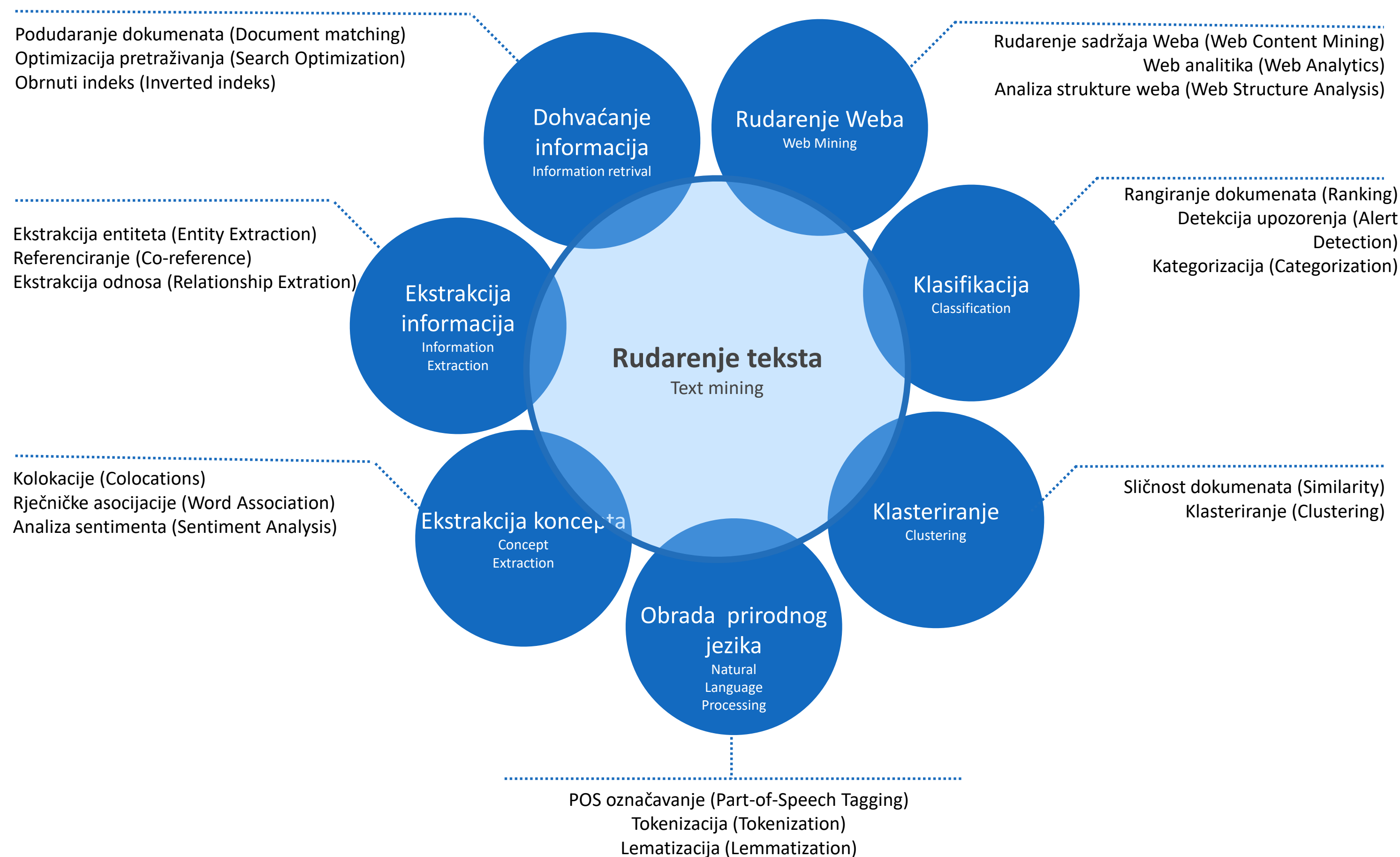
Zagreb, 10. svibnja 2019.

- 1 Uvod
- 2 Označivanje u računalnom modelu
- 3 Računalna realizacija modela
- 4 WOS/SOW strukture
- 5 Integracija u LOD oblak
- 6 Integracija s drugim vanjskim resursima (API)

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

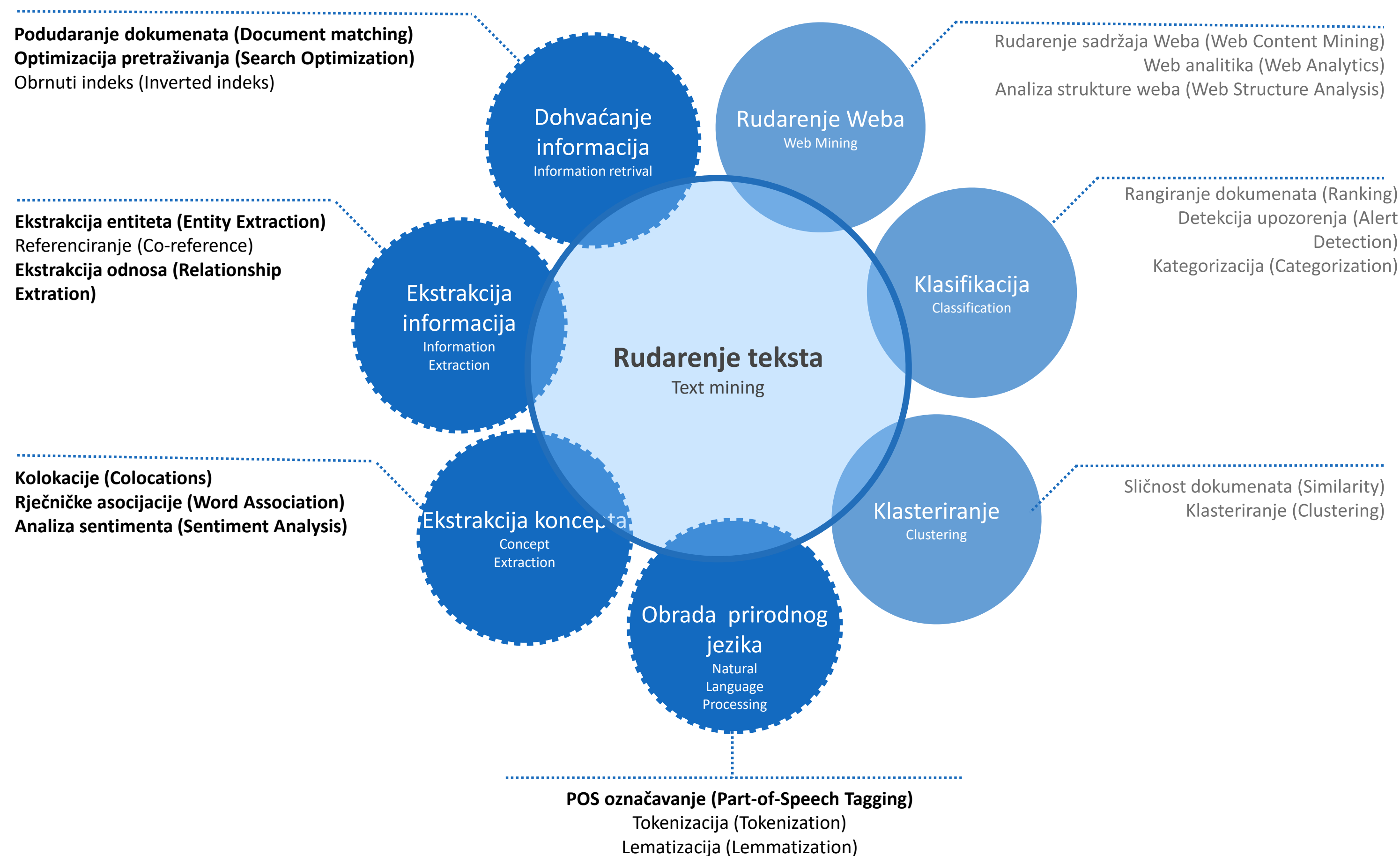


Rudarenje teksta – obrada prirodnog jezika



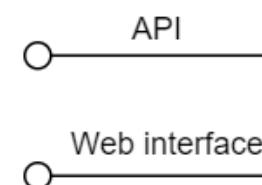
Stohastički vs. deterministički model

Rudarenje teksta – obrada prirodnog jezika



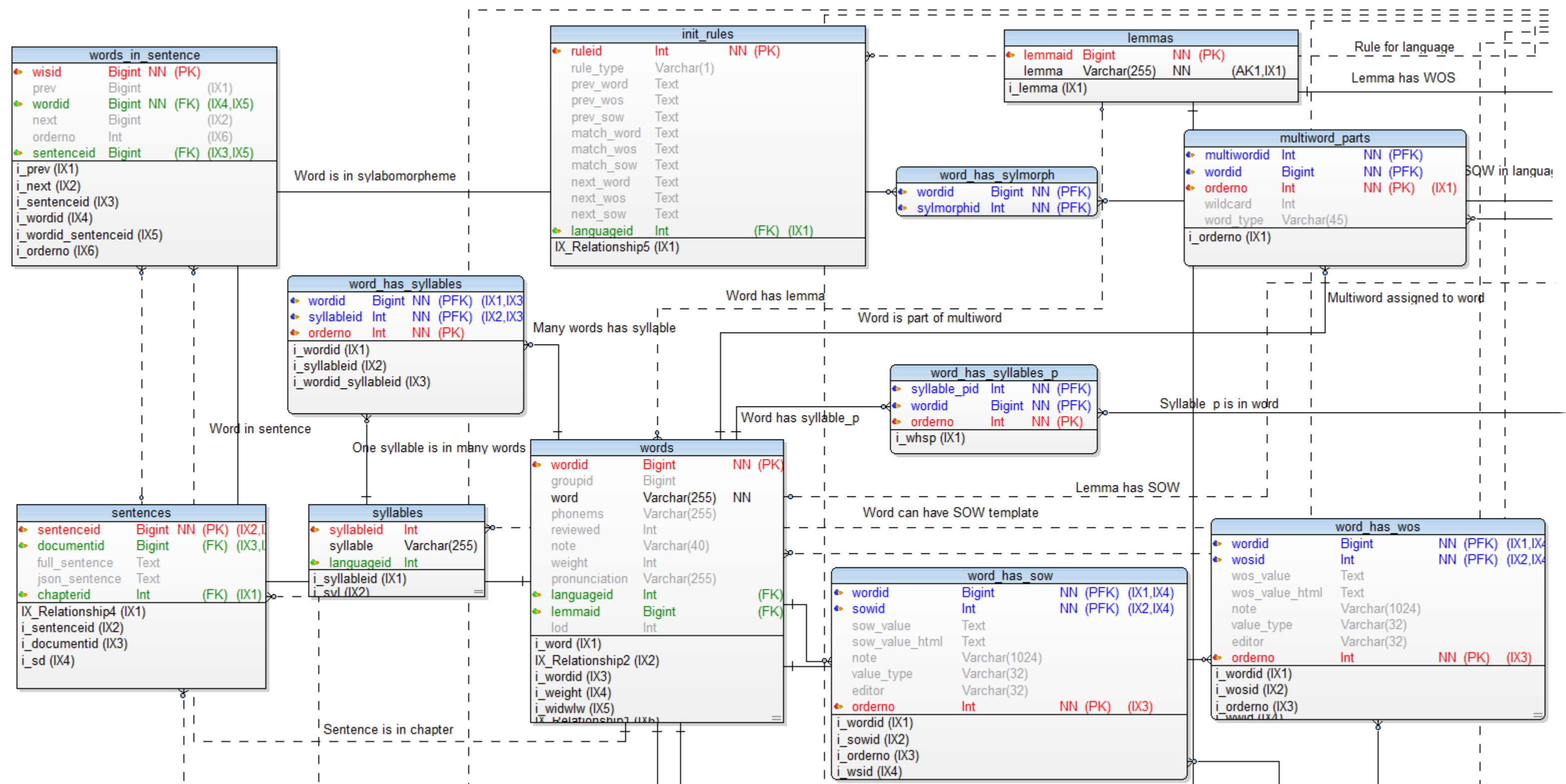
Stohastički vs. deterministički model

- 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

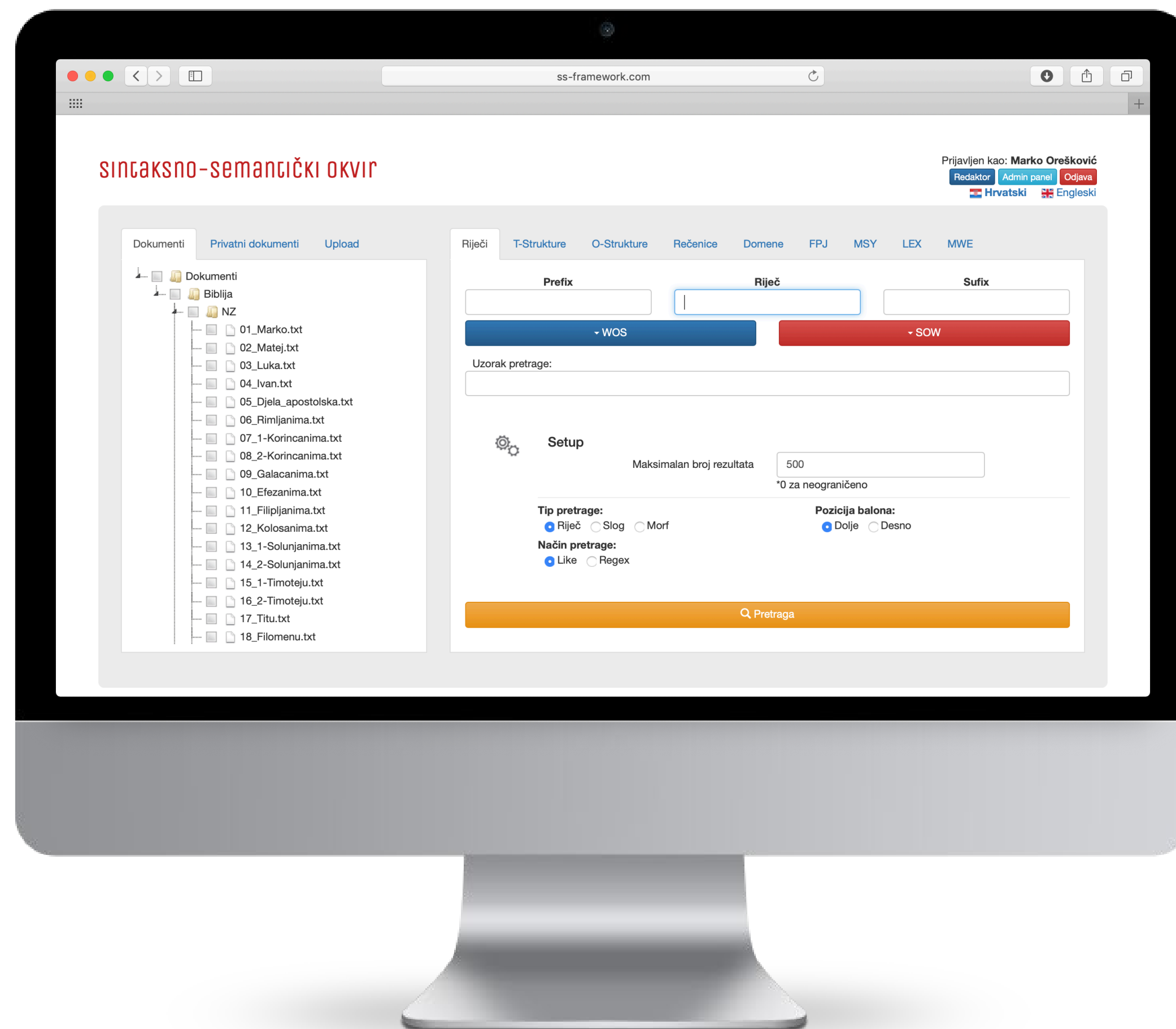


➤ Web aplikacija

➤ 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

- ▾ Vrsta riječi
 - Imenica
 - Zamjenica
 - Pridjev
 - Broj_vr
 - Glagol
 - Prilog
 - Prijedlog
 - Veznik
 - Uzvik
 - Čestica
 - Kratica
 - interpunkcija
 - Kombinabilne vrste
 - Akcentuacija
 - Rod
 - Broj
 - Padež

SOW

- ▾ Opće
 - Obilježje
 - Živo
 - Pojam
 - Tvar
 - Tvorevina
 - Relacija
 - Stanje
 - Proces
 - Prostor
 - Vrijeme
 - Terminološko
- ▾ Ime
 - Antroponim
 - Ime
 - Prezime
 - Nadimak
 - Toponim

➤ WOS – word of speech (gramatička obilježja)

➤ SOW – semantic of word (semantička obilježja)

Povezanost riječi s repozitorijima i enciklopedijom



WOS

Vrsta riječi

- Imenica
- Zamjenica
- Pridjev
- Broj_vr
- Glagol
- Prilog
- Prijedlog
- Veznik
- Uzvik
- Čestica
- Kratice
- Interpunkcija

Kombinabilne vrste

- Imenica
- Pridjev

A B C Č Ć D DŽ Đ E F G H I J K L LJ M N NJ O P Q R S Š T U V Z Ž X Y

NA ND NE NG NI NJ NM NO NT NU NÂ

NADA

Lema: nada

Slogovi: na-da

Morfovi: nada-

WOS: Vrsta riječi • Imenica Rod • Ženski Broj • Množina Padež • Genitiv

SOW: Stav • Pozitivno [2] Stav • Negativno [2] CroWN • Definicija [3] CroWN • Sinonim [2] CroWN • Antonim [4] CroWN • Hipernim [8] ENC • Definicija

NADA

Lema: nada

Slogovi: na-da

Morfovi: nada-

WOS: Vrsta riječi • Imenica Rod • Ženski Broj • Jednina Padež • Nominativ

SOW: Stav • Pozitivno [2] Stav • Negativno [2] CroWN • Definicija [3] CroWN • Sinonim [2]

NADA

Lema: nad

Slogovi: na-da

Morfovi: nada-

WOS: Vrsta riječi • Prijedlog • Uz A

SOW: CroWN • Definicija [3] CroWN • Sinonim [2] CroWN • Antonim [4] CroWN • Hipernim [8] ENC • Definicija

nada, očekivanje da će se ispuniti želja; očekivanje da će se ostvariti nešto što se želi kao dobro i kao ispunjavajuće za osobu. Zbog toga je u pravilu usmjerena na budućnost, pa katkad dolazi do zanemarivanja izravne zbiljnosti i do nekritičkoga vjerovanja u to da je ono što se očekuje bitnije i važnije od sadašnjega života, te nada u tom smislu prerasta u utopiju. S obzirom na ono transcendentno i apsolutno, nada postaje izvorom religioznih vjerovanja. U filozofijskom smislu, nada je akt duhovnoga doživljavanja kojim se otkrivaju dubine duše (G. Marcel), temeljni način iskušavanja još-ne-bitka (E. Bloch) koji u tom činu postaje smislom našega sadašnjega bitka u svijetu, odnosno mišljenja i djelovanja na niti vodilji tako razumljene nade.

- 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)

AP

Riječi:

ap-strak-ci-ja ap-strak-tna ap-strak-tno ap-surd ap-sur

ap-sti-ni-ra-ti di-ap-si-da di-ap-sid-ni ap-so-lu-ti-sti-e-k

ap-sti-ni-ra-nje ap-sti-nen-ci-ja ap-si-dom mo-ap-skom

mo-ap-sku mo-ap-sko-ga mo-ap-ski mo-ap-skim mo-a

mo-ap-ce mo-ap-ci mo-ap-skih mo-ap-ci-ma mo-ap-sl

mo-ap-ka mo-ap-ku mo-ap-sko-me mo-ap-ke ap-sa-lo

ap-ci-ha ap-si-da ap-si-da-ma ap-si-de ap-si-di ap-

ap-so-lu-tan ap-so-lu-ti-zam ap-so-lu-ti-za-ma ap-so-lu-

ap-so-lu-ti-zi-ra-hu ap-so-lu-ti-zi-ra-raj ap-so-lu-ti-zi-raj-mo

ap-so-lu-ti-zi-ra-ju ap-so-lu-ti-zi-ra-la ap-so-lu-ti-zi-ra-le

ap-so-lu-ti-zi-ra-lo ap-so-lu-ti-zi-ram ap-so-lu-ti-zi-ra-mo

ap-so-lu-ti-zi-ra-smo ap-so-lu-ti-zi-ra-ste ap-so-lu-ti-zi-ra-

ap-so-lu-ti-zi-rav-ši ap-so-lu-ti-zi-raš ap-so-lu-ti-zi-ra-še

WOS

Vrsta riječi

- ☐ Imenica
- ☐ Zamjenica
- ☐ Pridjev
- ☐ Broj_vr
- ☐ Glagol
- ☐ Prilog
- ☐ Prijedlog
- ☐ Veznik
- ☐ Uzvik
- ☐ Čestica
- ☐ Kratica
- ☐ Interpunkcija

Kombinabilne vrste

- ☐ Imenica
- ☐ Pridjev
- ☐ Prilog

A B C Č Ć D DŽ Đ E F G H I J K L LJ M N

Postoji 4042 riječi koje zadovoljavaju kriterij pretrage. P

LABAV ČVOR

Elementi: labav čvor

Vrsta: Jednostavna višerječnica

WOS:

SOW:

LABAV REŽIM

Elementi: labav režim

Vrsta: Kolokacija Dr. Stefan Rittgasser

WOS:

SOW: Kolokacija

LABAVA CARINSKA UNIJA

Elementi: labava carinska unija

Vrsta: Kolokacija Dr. Stefan Rittgasser

WOS:

SOW: Kolokacija

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



➤ Riješeni kompleksni problemi višeznačnosti

Words:	Marija	usta	i	pođe	u	gorje	.
WOS:	Imenica	Glagol	Veznik	Glagol	Prijedlog	Imenica	Interpunktija
	Nominativ			Infinitiv		Nominativ	
	Jednina					Jednina	
SOW:	Osobno ime						

Words:	Marijina	usta	i	oči	su	govorile	.
WOS:	Pridjev	Imenica	Veznik	Imenica	Pomoćni glagol biti	Glagol	Interpunktija
	Nominativ	Nominativ					
	Jednina	Množina					
SOW:	Osobno ime						

Izvlačenje sintaktičko-semantičke informacije iz teksta



Sentence 1:

Ona mu je majka.

11_1-Kraljevima.txt

ona	mu	je	majka	.
WOS: Vrsta riječi • Zamjenica • Pokazna Rod • Ženski Broj • Jednina Padež • Nominativ Naglašenost • Nenaglašen	WOS: Vrsta riječi • Zamjenica • Osobna Rod • Muški Broj • Jednina	WOS: Vrsta riječi • Glagol • Pomoćni • Biti Broj • Jednina Naglašenost • Nenaglašen Osoba • 3. Vid • Nesvršen Vrijeme • Present	WOS: Vrsta riječi • Imenica Rod • Ženski Broj • Množina Padež • Genitiv SOW: .OWL • owl:sameAs [2] BabelNet • Definicija BabelNet • Kategorija [3] CroWN • Definicija CroWN • Sinonim [4] CroWN • Antonim [2] CroWN • Hipernim	WOS: Vrsta riječi • interpunkcija • Svršetak • .

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

Označivanjem do umrežene informacije

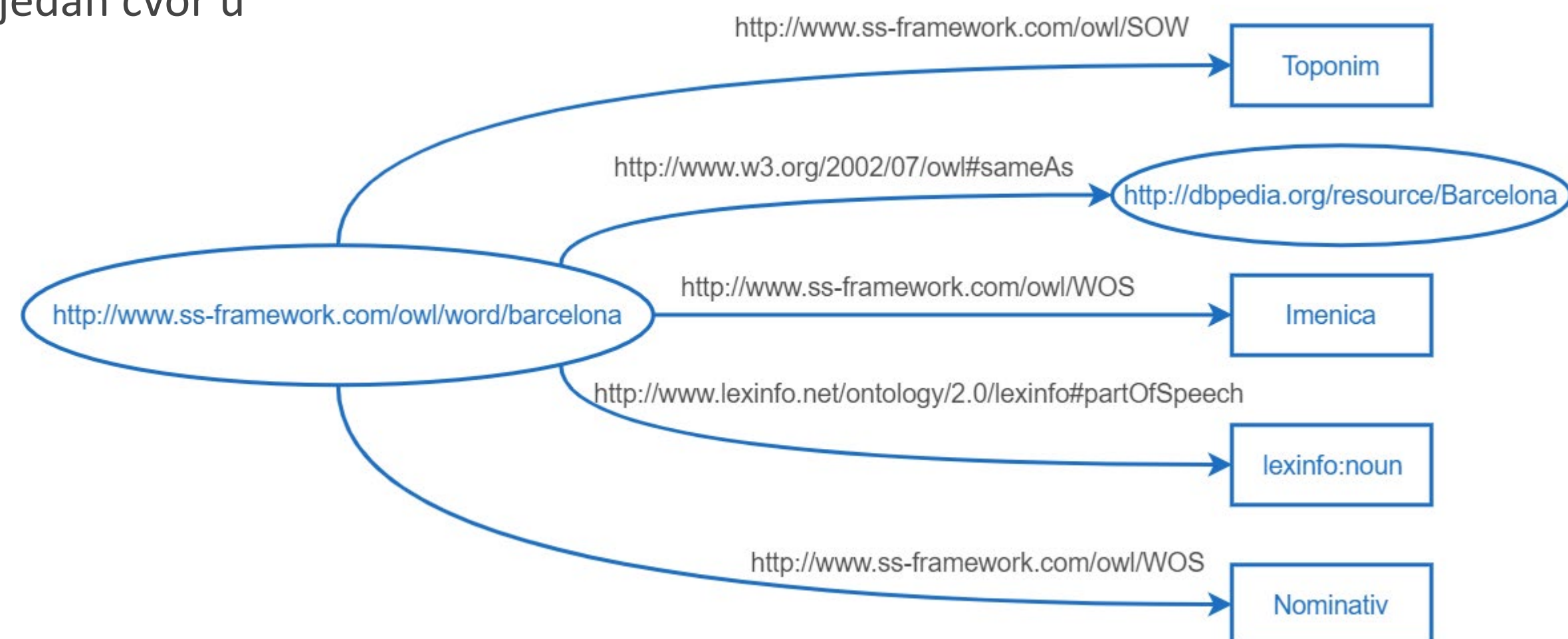


- Poseban tip SOW tagova **owl:sameAs** za povezivanje s vanjskim resursima
- Drugi WOS/SOW tagovi transformiraju se u RDF trojke

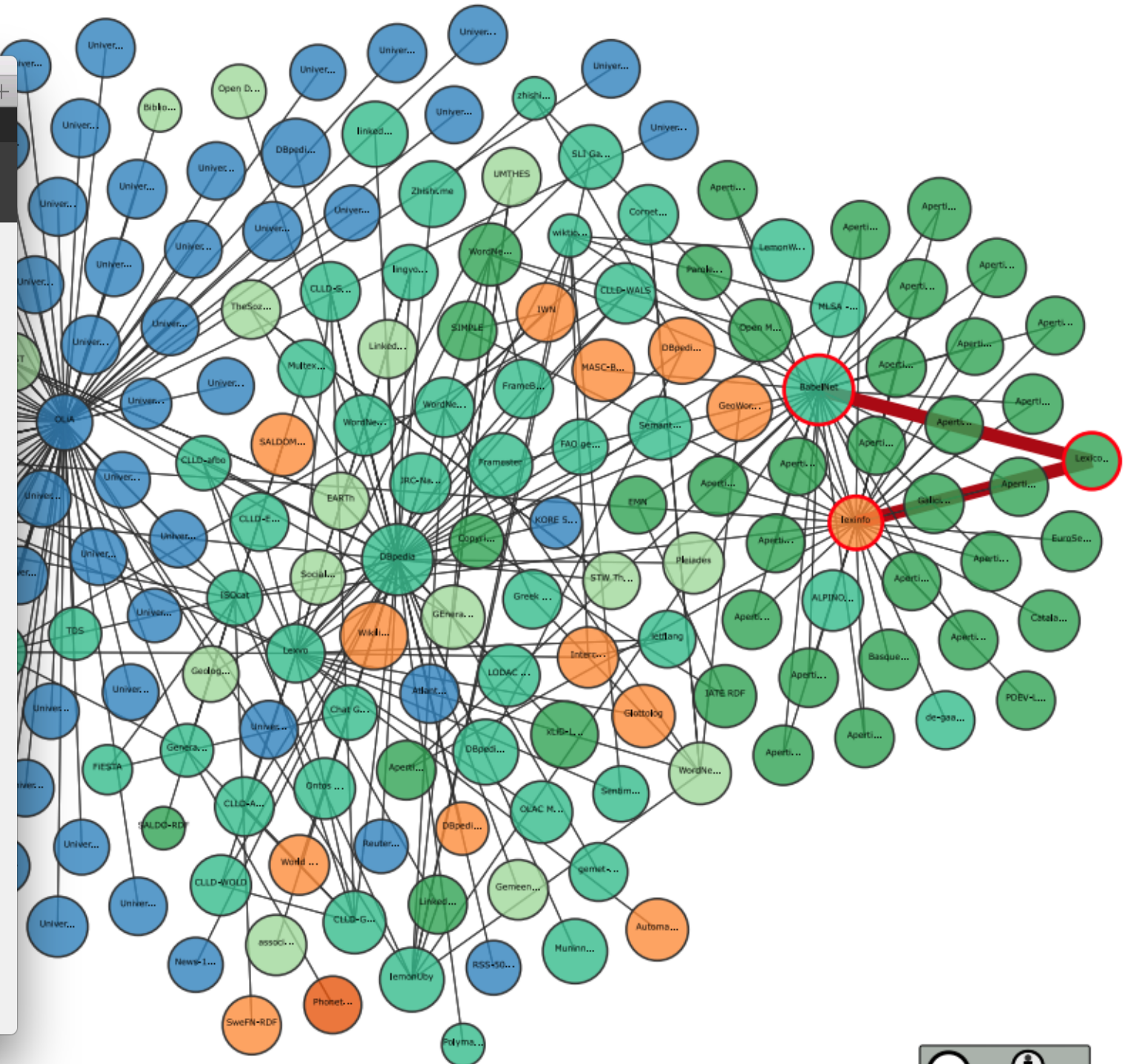
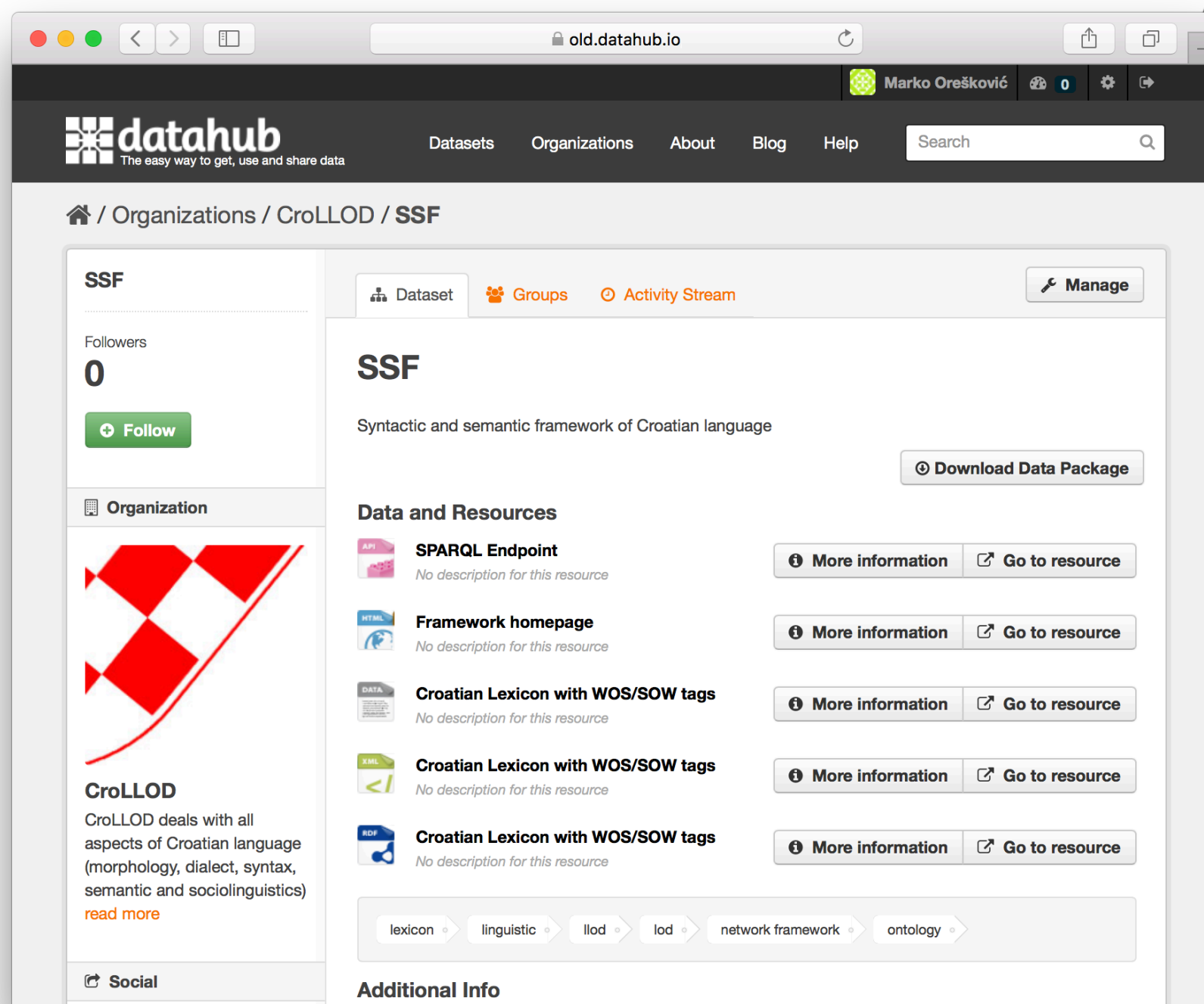
BARCELONA
Lemma: barcelona
Slogovi: bar-ce-lo-na
WOS: Vrsta riječi • Imenica Rod • Ženski Broj • Jednina Padež • Nominativ
SOW: OWL • owl:sameAs Ime • Toponim

<http://dbpedia.org/resource/Barcelona>

- 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)



- ④ Integracija s vanjskim resursima
- ④ REST API preko HTTP, primjer Python koda:

```
1  # -*- coding: utf-8 -*-
2  import requests
3
4  # API key
5  key = "429ae4bfe8088f071abef86ac021653b"
6
7  # Execute in Python, Haskell, SPARQL, R
8  program = "Python"
9
10 # Code to be executed
11 code = "=ChangeTense('vidim plavu kuću', 'aorist')"
12 url = "http://www.ss-framework.com/api/fpj"
13
14 req = requests.post(url, data = {'apiKey':key, 'program':program, 'code':code})
15
16 # Set the output encoding to UTF-8
17 req.encoding = 'UTF-8'
18
19 # Print the output
20 print req.text
```

④ Odgovor: {"status":200, "result":"vidjeh plavu kuću"}

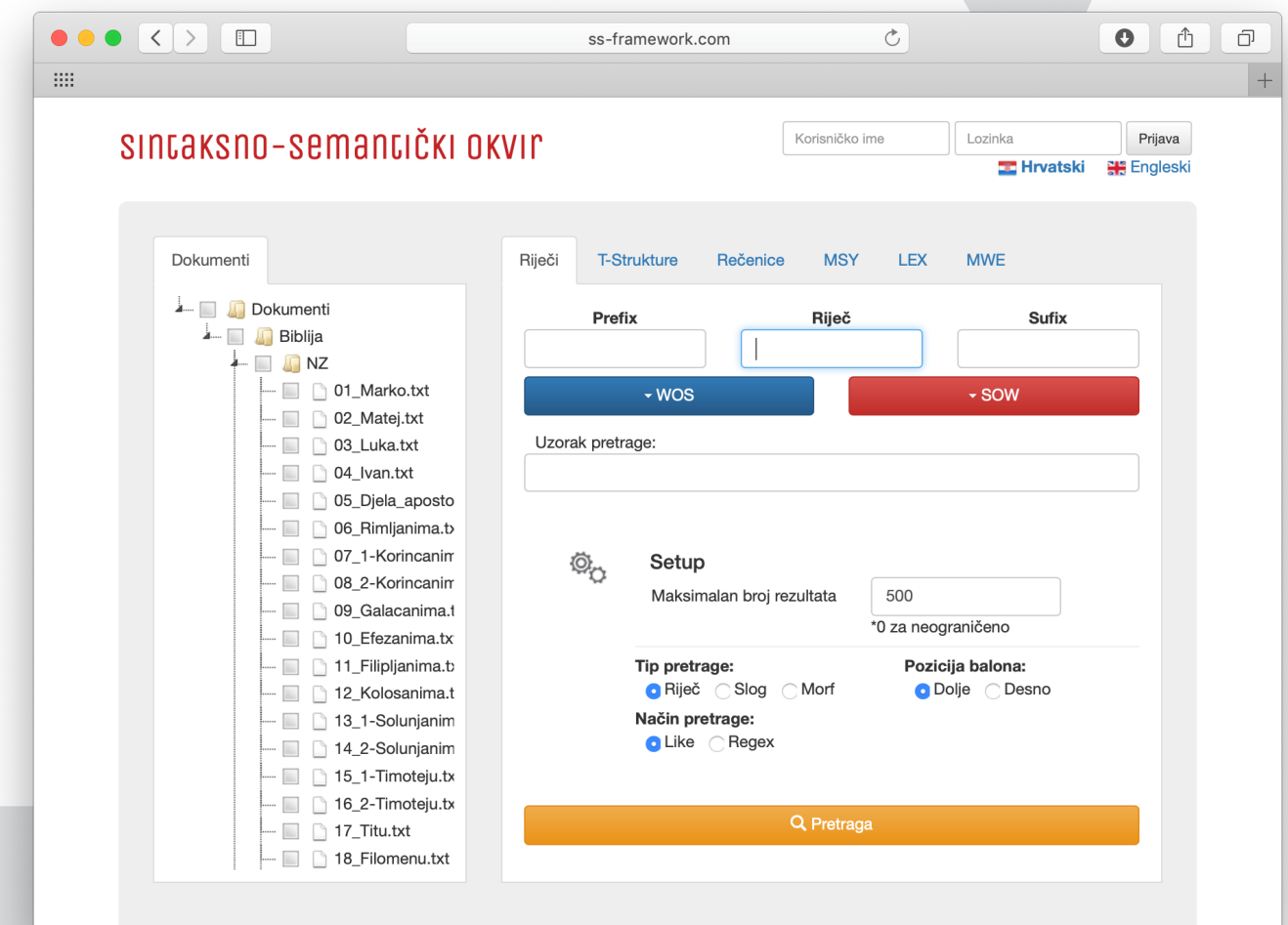
Application programming interface (API)



➤ 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.