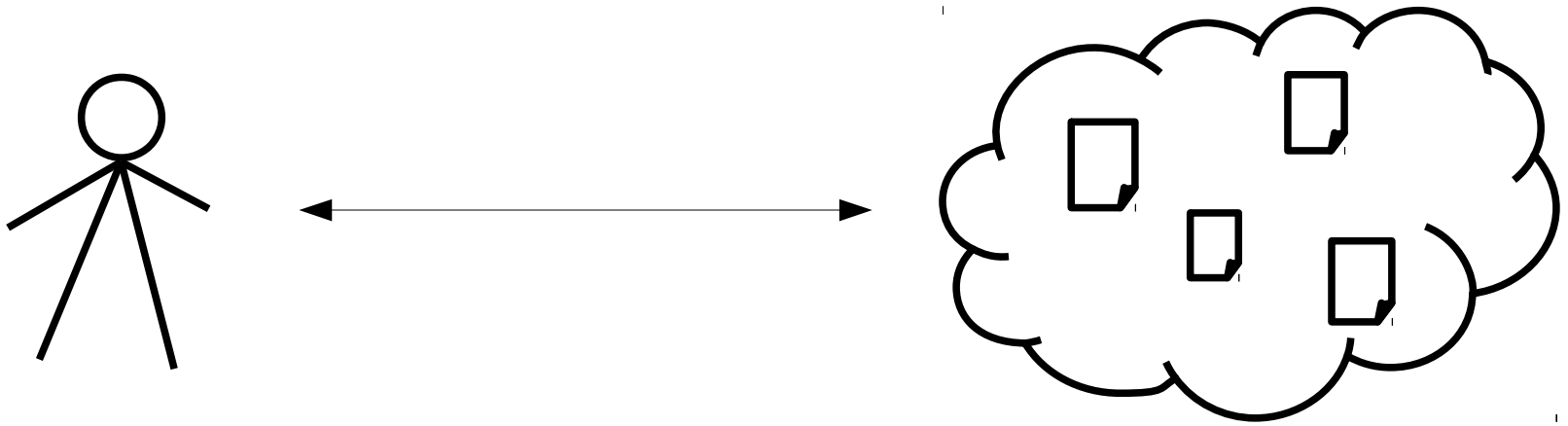
Information Extraction and Natural Language Processing

Marián Šimko

25. 11. 2015

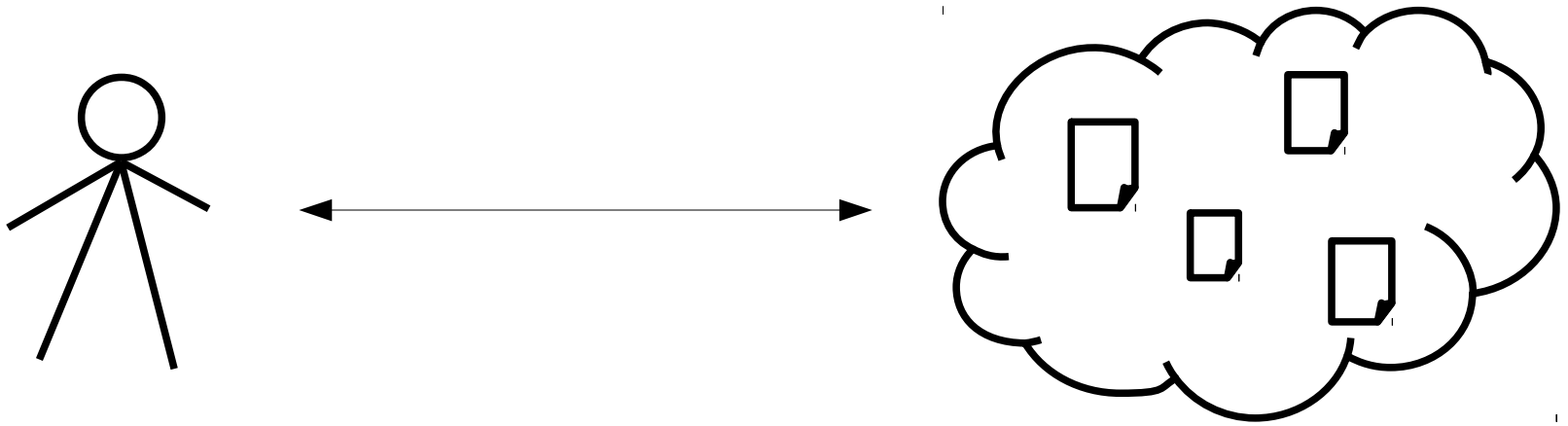
Prístup k informáciám

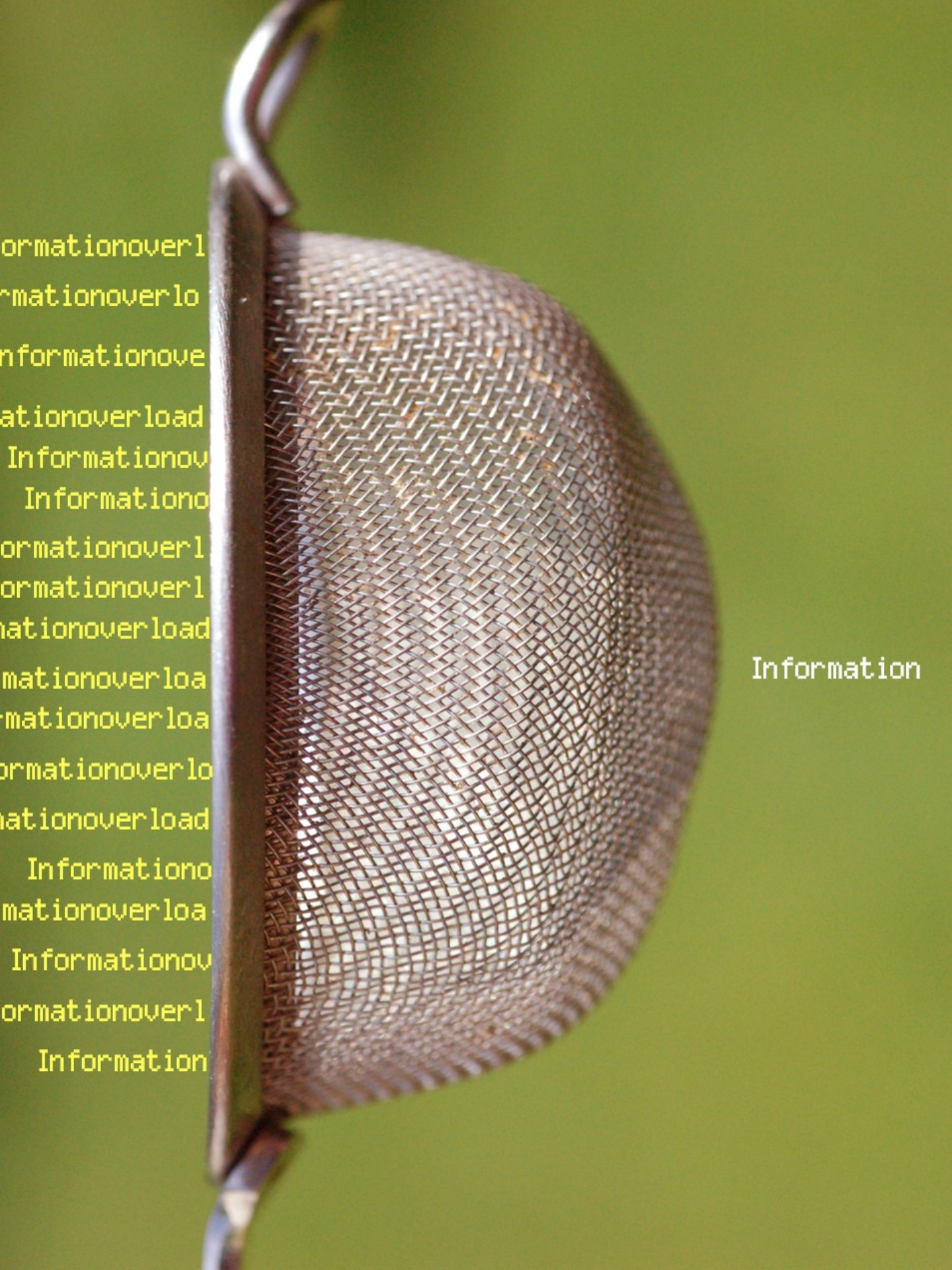


Prístup k informáciám

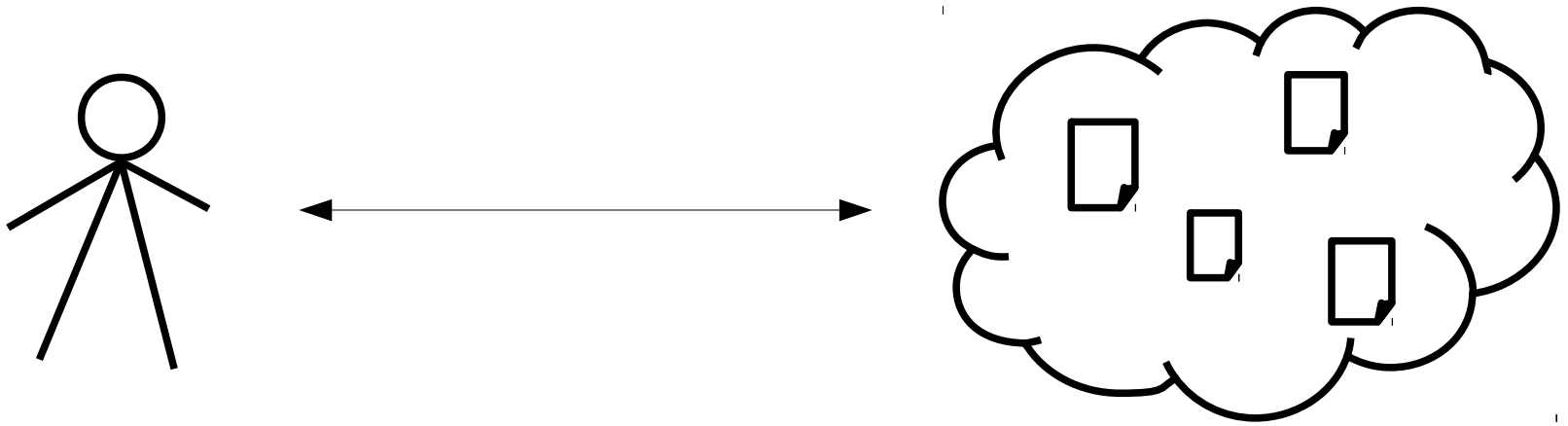
November 2015:

55 miliárd webových stránok

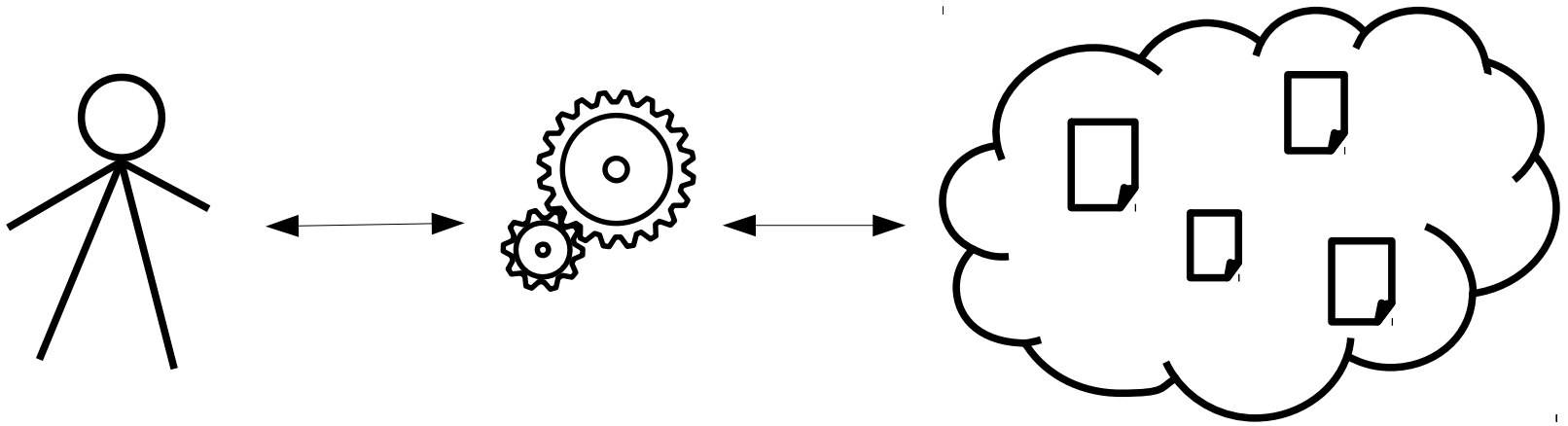



ormationoverl
rmationoverlo
nformationove
ationoverload
Informationov
Informationo
ormationoverl
ormationoverl
nationoverload
mationoverloa
rmationoverloa
ormationoverlo
nationoverload
Informationo
mationoverloa
Informationov
ormationoverl
Information
Information

Prístup k informáciám



Prístup k informáciám



Naše informačné potreby

- Information retrieval
- Information extraction
- Document classification/categorisation
- Reasoning
- ...
- Artificial intelligence


INFORMATION
KNOWLEDGE
RECOMMENDATION
TECHNOLOGY
DOSSIER
PERCEPTION
WORD
LEARNING
INTERNET
DOCUMENT
DATA
NEWS
DIRECTIVE
WISDOM
ADVICE
FACTS
SYMBOL
MESSAGE
SOURCE
DIRECTIONS
TEACH
MODERN
BOOKS
STUDY
INFO
SEARCH
WEB
MEDIA
GUIDANCE
CONSCIOUSNESS
PERCEPTION
123RF

Informácie –

kódované v prirodzenom jazyku

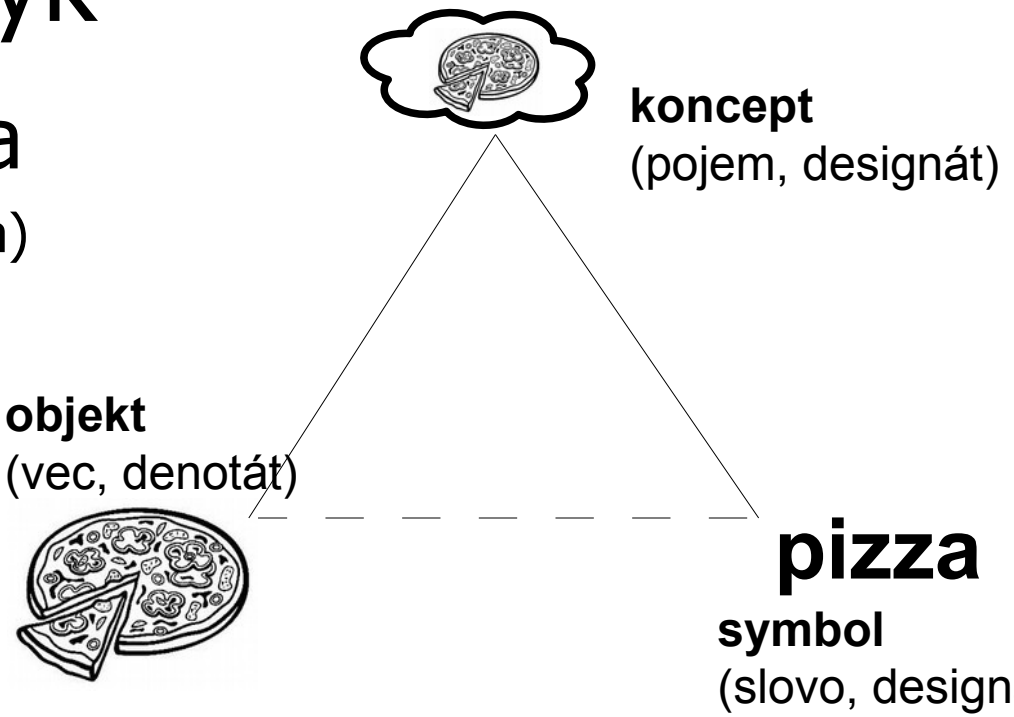
Informácie sú v dokumentoch.

Dokumenty =

Prirodzený jazyk

Písomná podoba

(symbolická reprezentácia)



Prirodzený jazyk =
komunikačný prostriedok

Komunikačný postriedok?

Umelá komunikácia

- Vysielateľ – prijímateľ
- Dáta
- Kompresia
- Samoopravný kód
  (napr. Hammingov kód)

Komunikačný postriedok?

Umelá komunikácia

- Vysielateľ – prijímateľ
- Dáta
- Kompresia
- Samoopravný kód (napr. Hammingov kód)

Prirodzený jazyk

- Autor – čitateľ
- Text
- Stratová kompresia
- Samoopravný kód (napr. gramatika)

Natural language processing

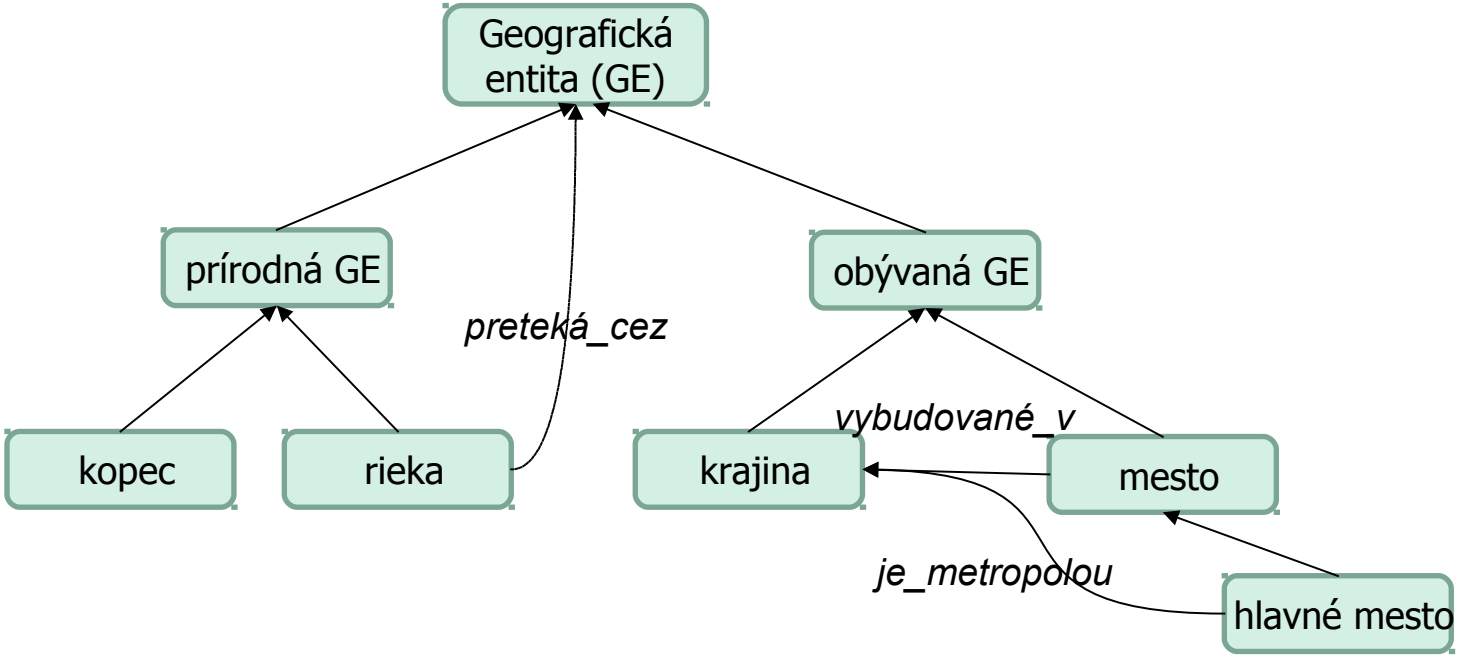
= spracovanie prirodzeného jazyka

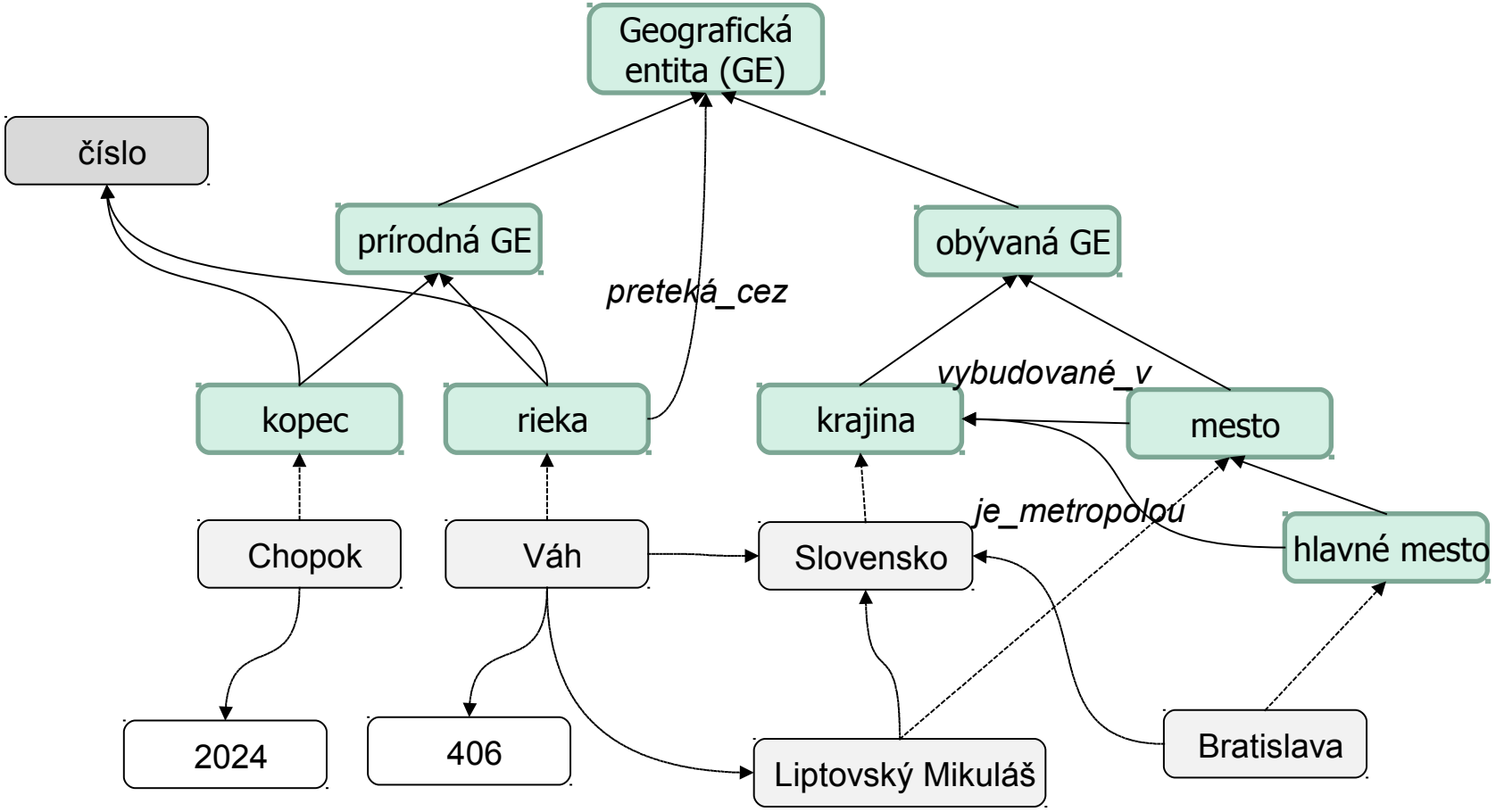
Nachádzanie informácií v komunikácii v prirodzenom jazyku

Príklad:

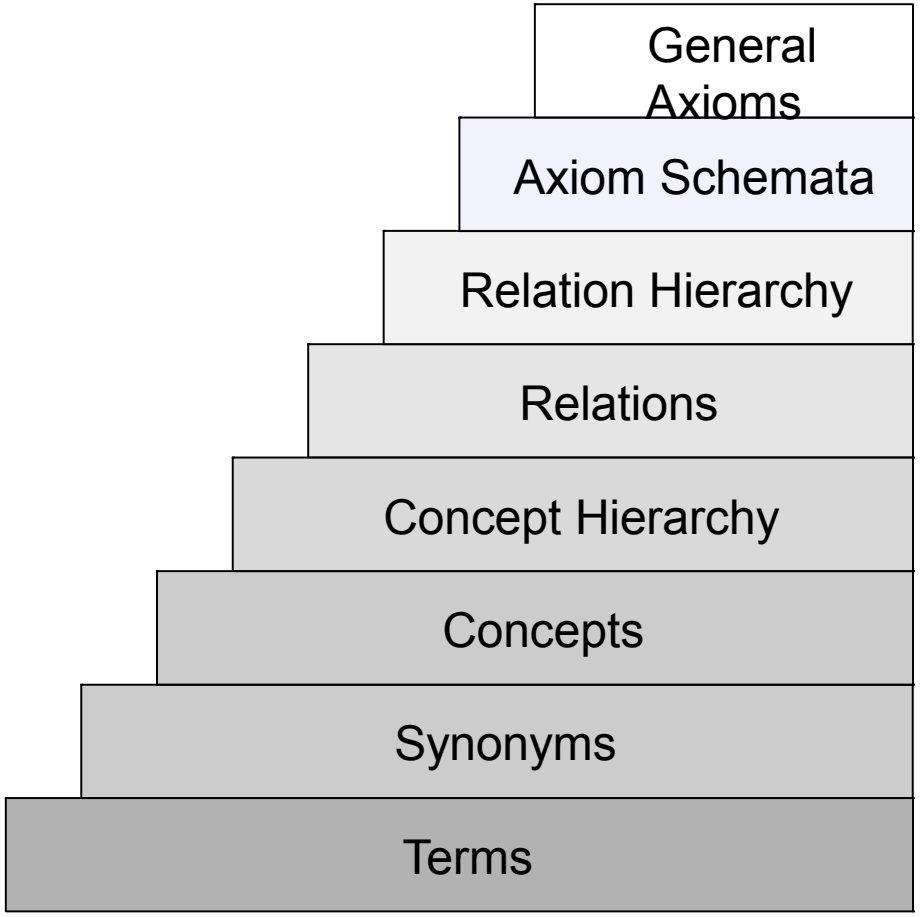
korpus dokumentov z geografie

Chceme opísať danú oblasť


Geografická entita (GE)
prírodná GE
obývaná GE
preteká_cez
kopec
rieka
vybudované_v
krajina
mesto
je_metropolou
hlavné mesto

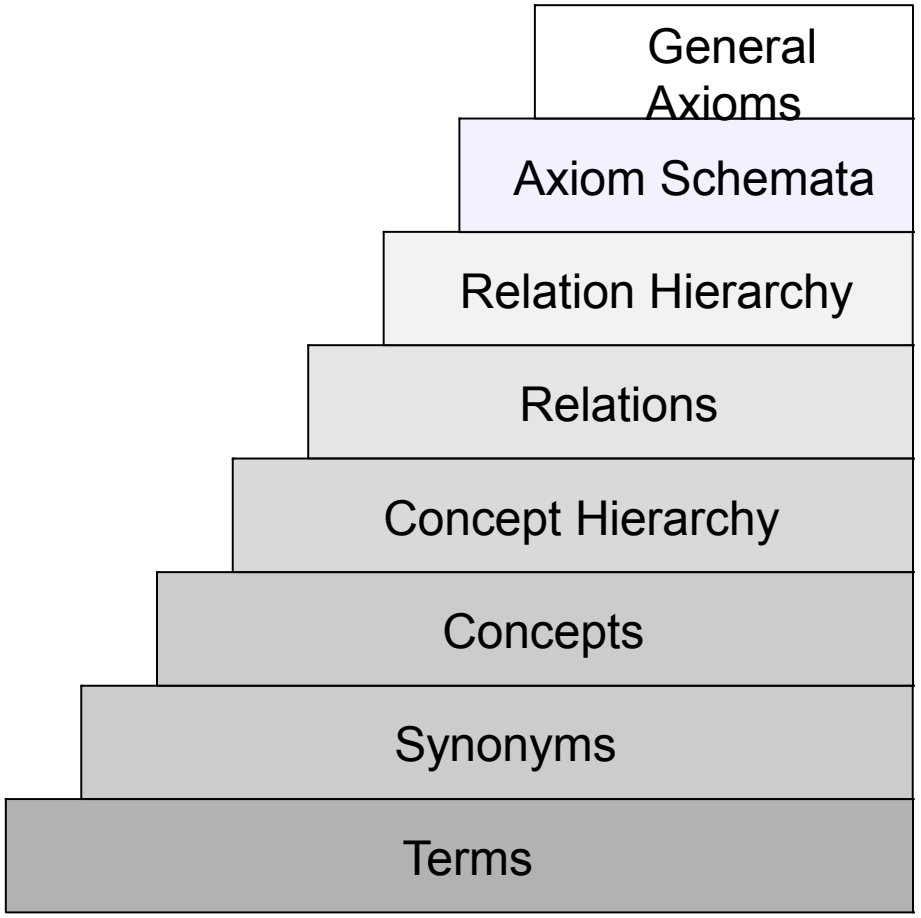

Geografická entita (GE)
číslo
prírodná GE
obývaná GE
preteká_cez
vybudované_v
kopec
rieka
krajina
mesto
je_metropolou
Chopok
Váh
Slovensko
hlavné mesto
2024
406
Liptovský Mikuláš
Bratislava

Ontológia



(Cimiano, 2006)

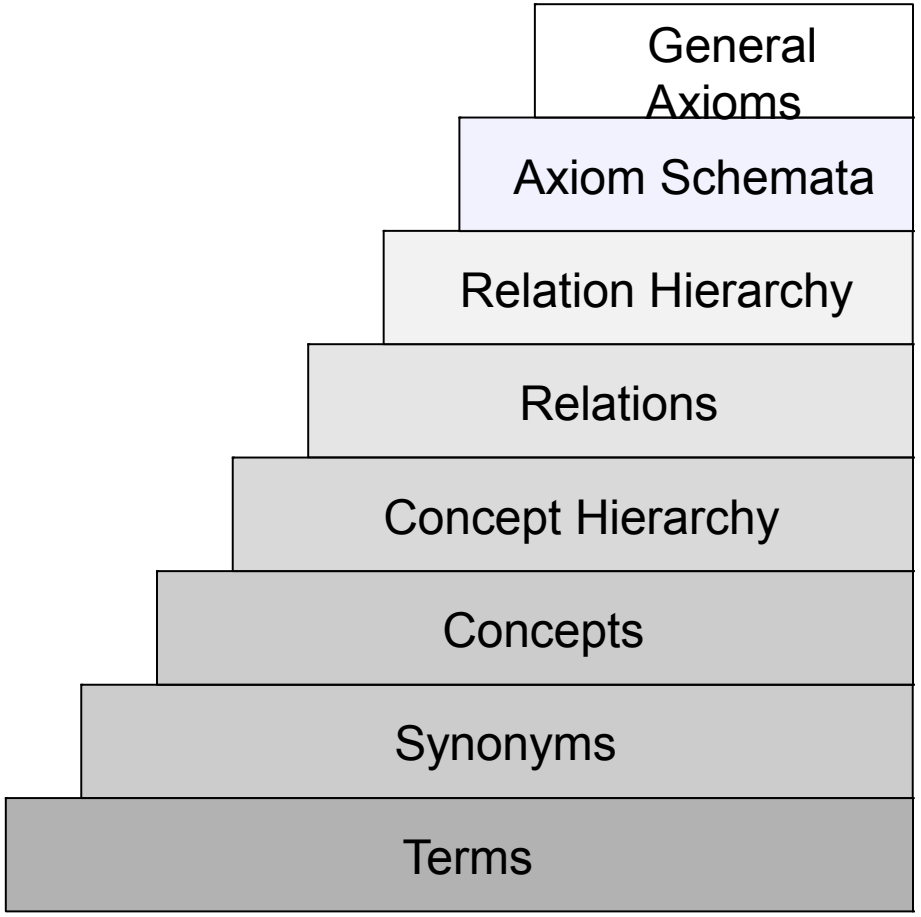
Ontológia



river, coutry, nation, city, capital,...

(Cimiano, 2006)

Ontológia

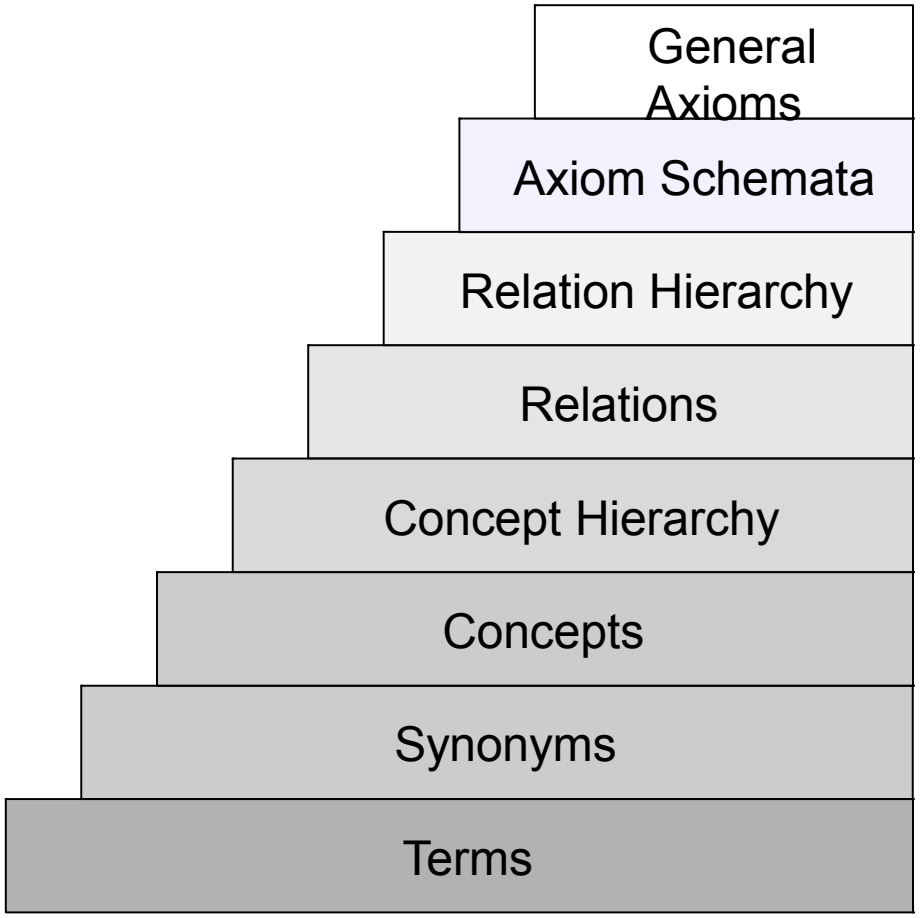


{country, nation}

river, coutry, nation, city, capital,...

(Cimiano, 2006)

Ontológia



c:= country := <i(c), ||c||, $Ref_C(c)$>

{country, nation}

river, coutry, nation, city, capital,...

(Cimiano, 2006)

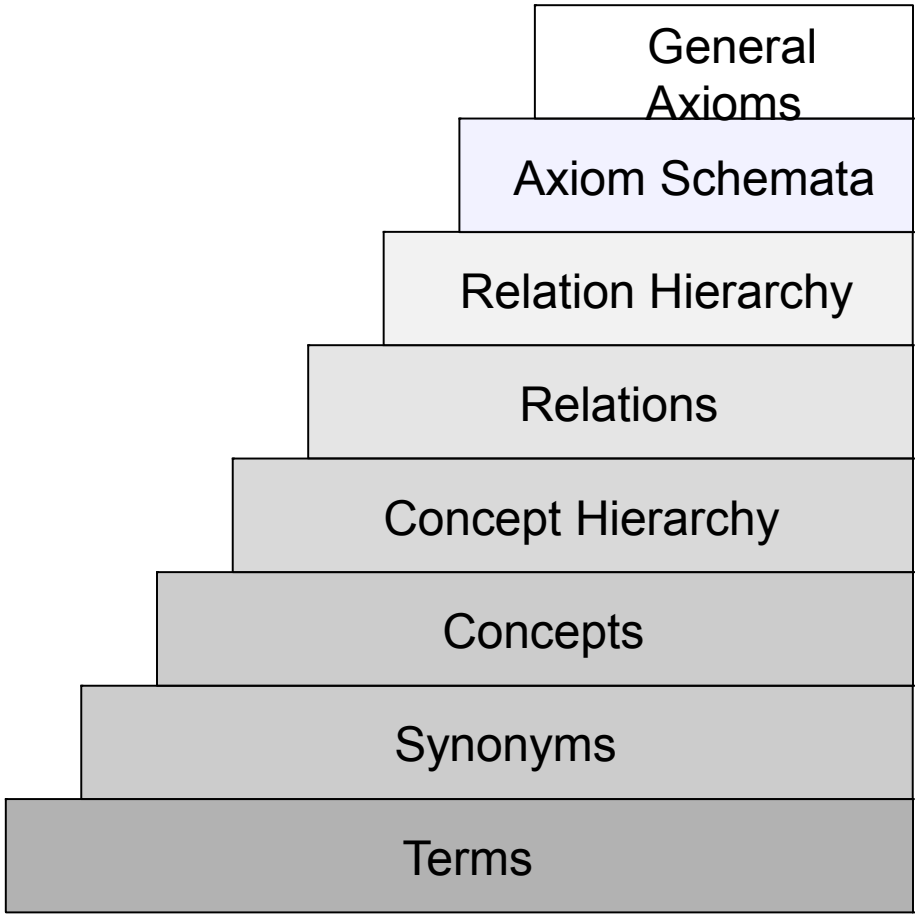
Ontológia

capital $\leq_C$ city, city $\leq_C$ GE

c:= country := <i(c), ||c||, Ref_C(c)>

{country, nation}

river, coutry, nation, city, capital,...



(Cimiano, 2006)

Ontológia

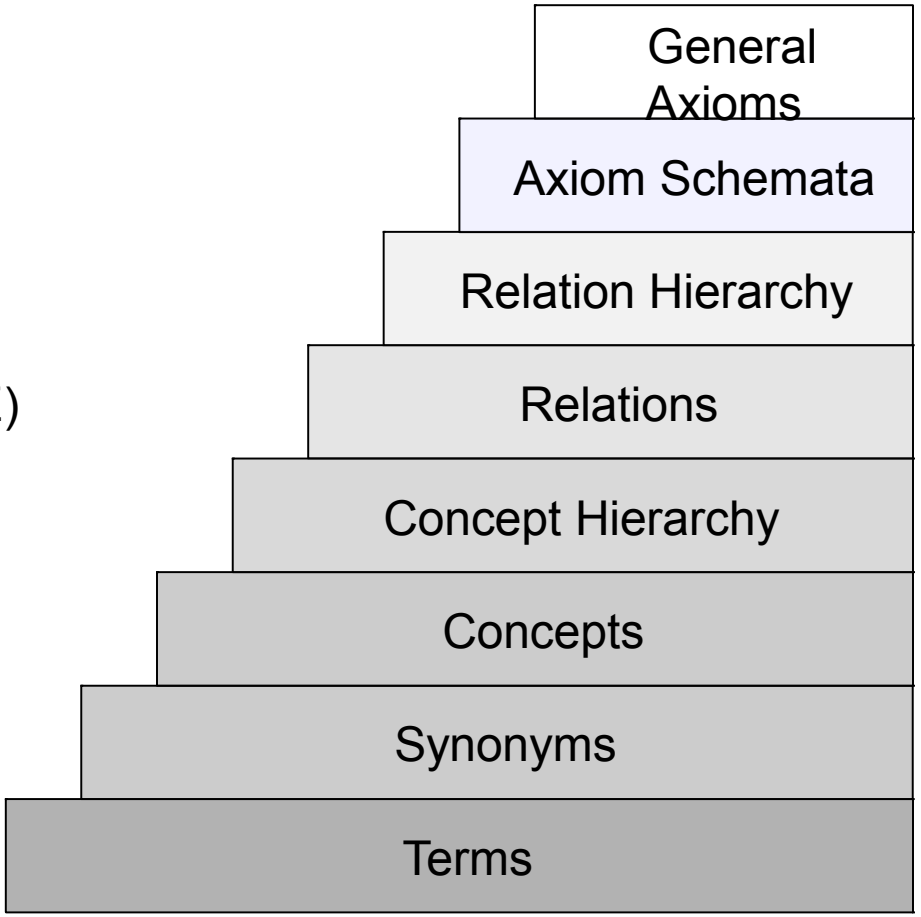
flow_through(dom:river, range:GE)

capital $\leq_c$ city, city $\leq_c$ GE

c:= country := <i(c), ||c||, $Ref_c(c)$>

{country, nation}

river, coutry, nation, city, capital,...



(Cimiano, 2006)

Ontológia

capital_of $\leq_R$ located_in

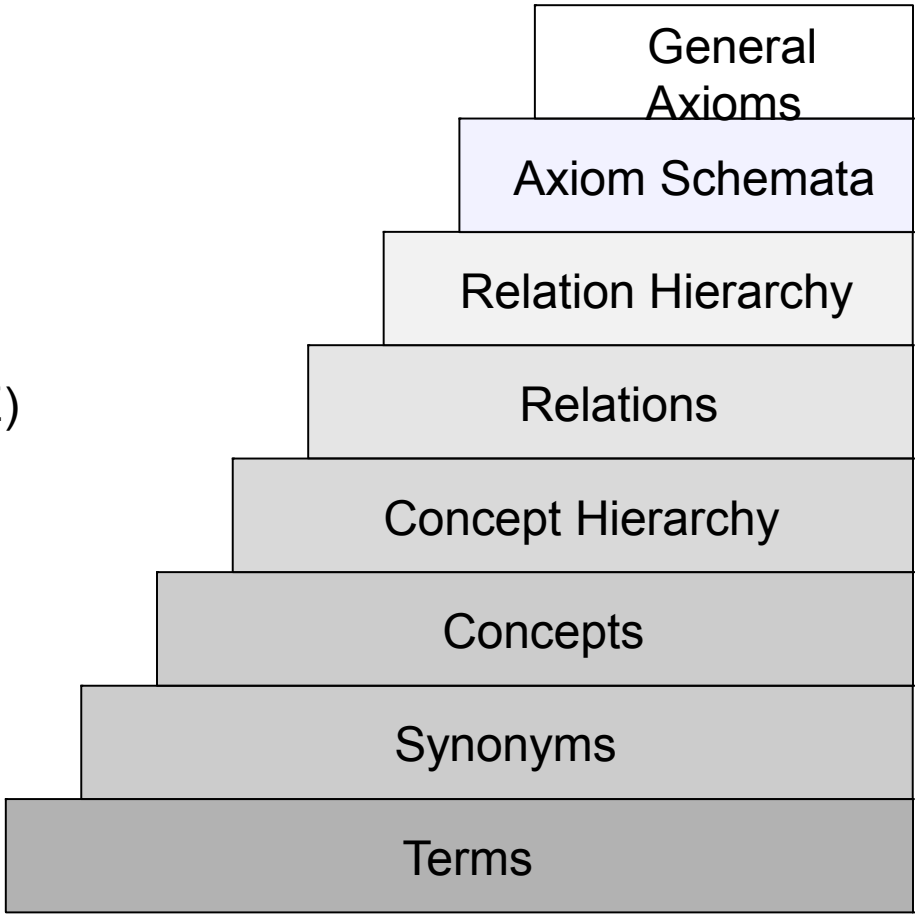
flow_through(dom:river, range:GE)

capital $\leq_C$ city, city $\leq_C$ GE

c:= country := <i(c), ||c||, $Ref_C(c)$>

{country, nation}

river, coutry, nation, city, capital,...



(Cimiano, 2006)

Ontológia

disjoint(river, mountain)

capital_of $\leq_R$ located_in

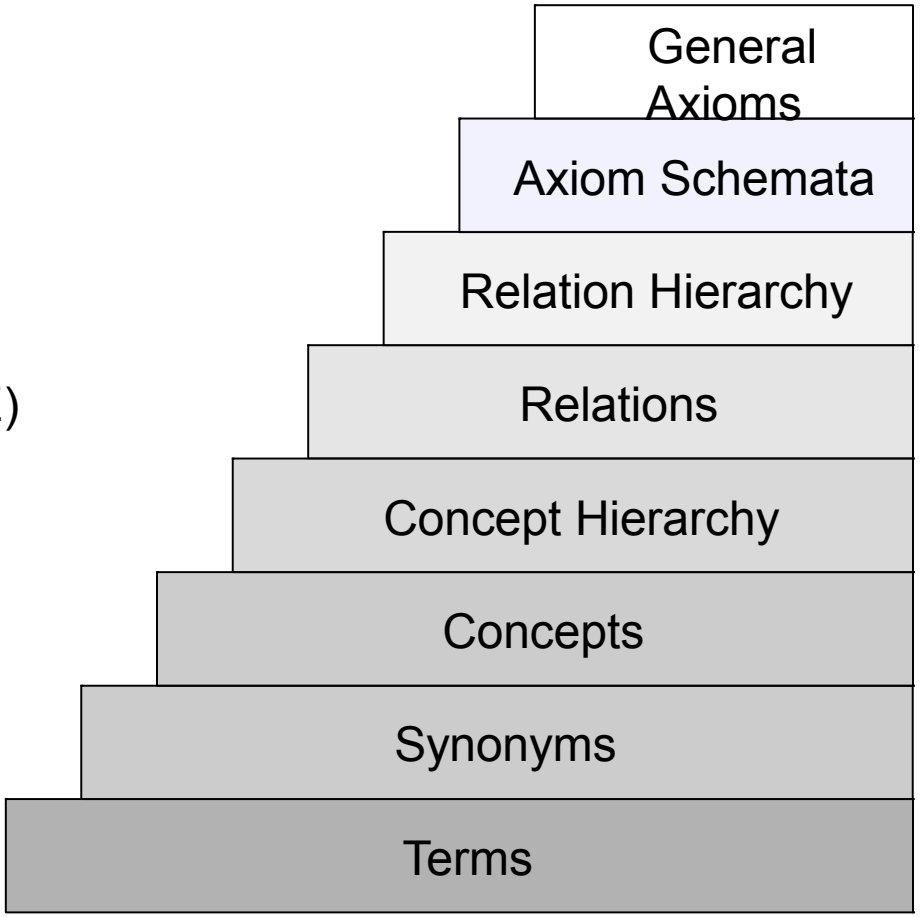
flow_through(dom:river, range:GE)

capital $\leq_C$ city, city $\leq_C$ GE

c:= country := <i(c), ||c||, $Ref_C(c)$>

{country, nation}

river, coutry, nation, city, capital,...



(Cimiano, 2006)

Ontológia

$\forall x(country(x) \rightarrow \exists y\ capital\_of(y,x) \wedge \forall z(capital\_of(z,x) \rightarrow y=z))$

disjoint(river, mountain)

capital_of $\leq_R$ located_in

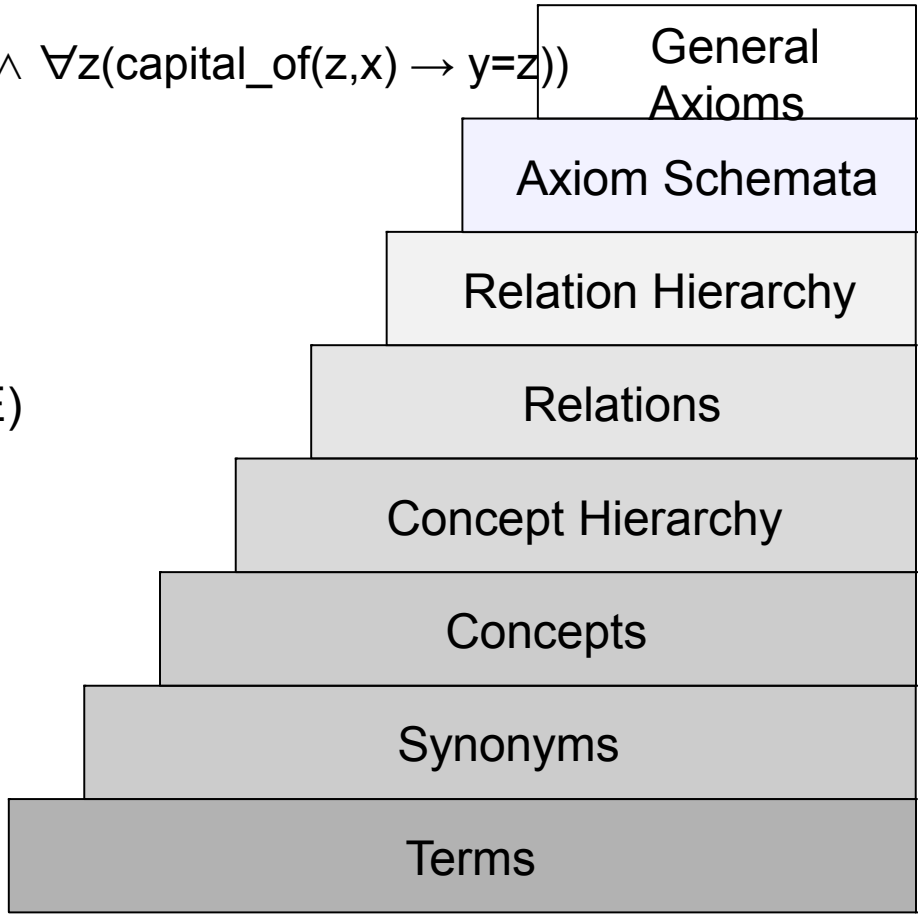
flow_through(dom:river, range:GE)

capital $\leq_C$ city, city $\leq_C$ GE

c:= country := <i(c), ||c||, $Ref_C(c)$>

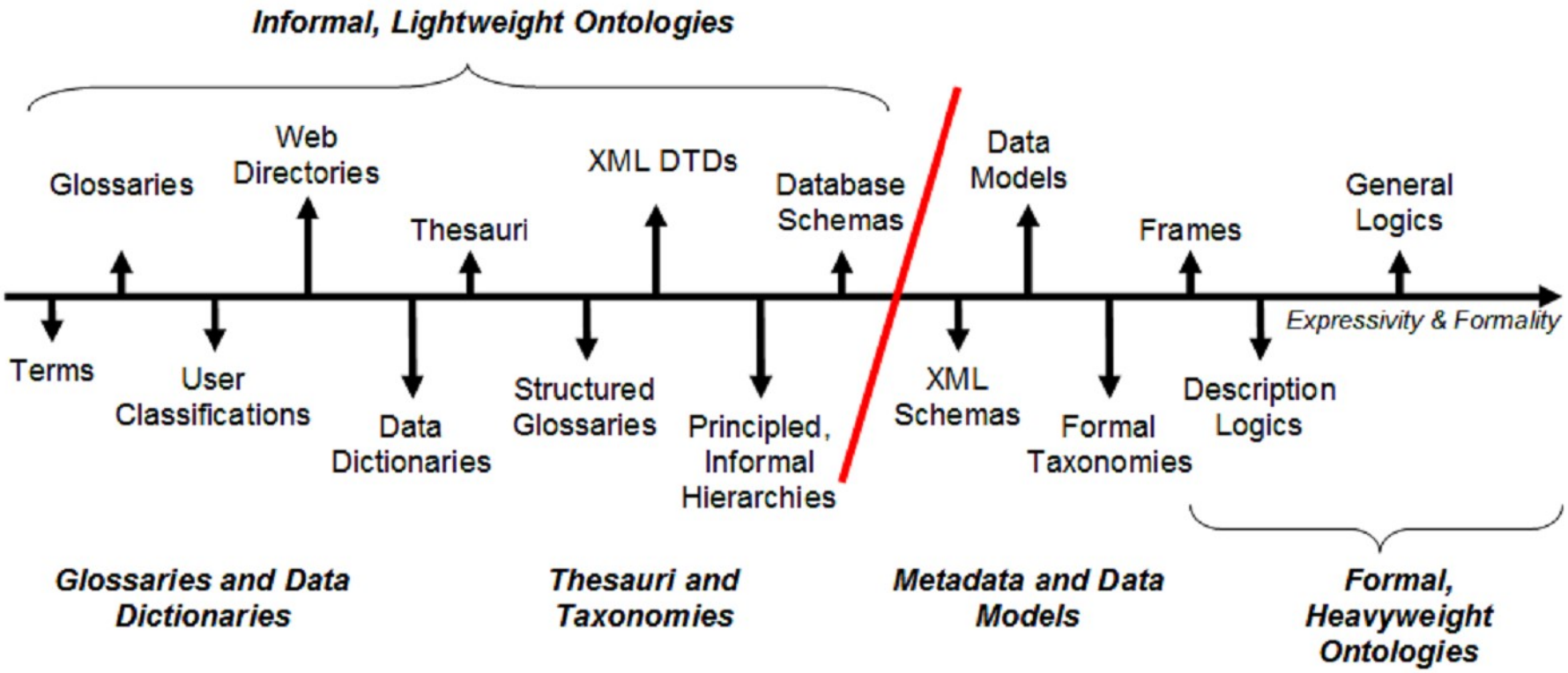
{country, nation}

river, coutry, nation, city, capital,...



(Cimiano, 2006)

Lightweight vs. Heavyweight ontológie

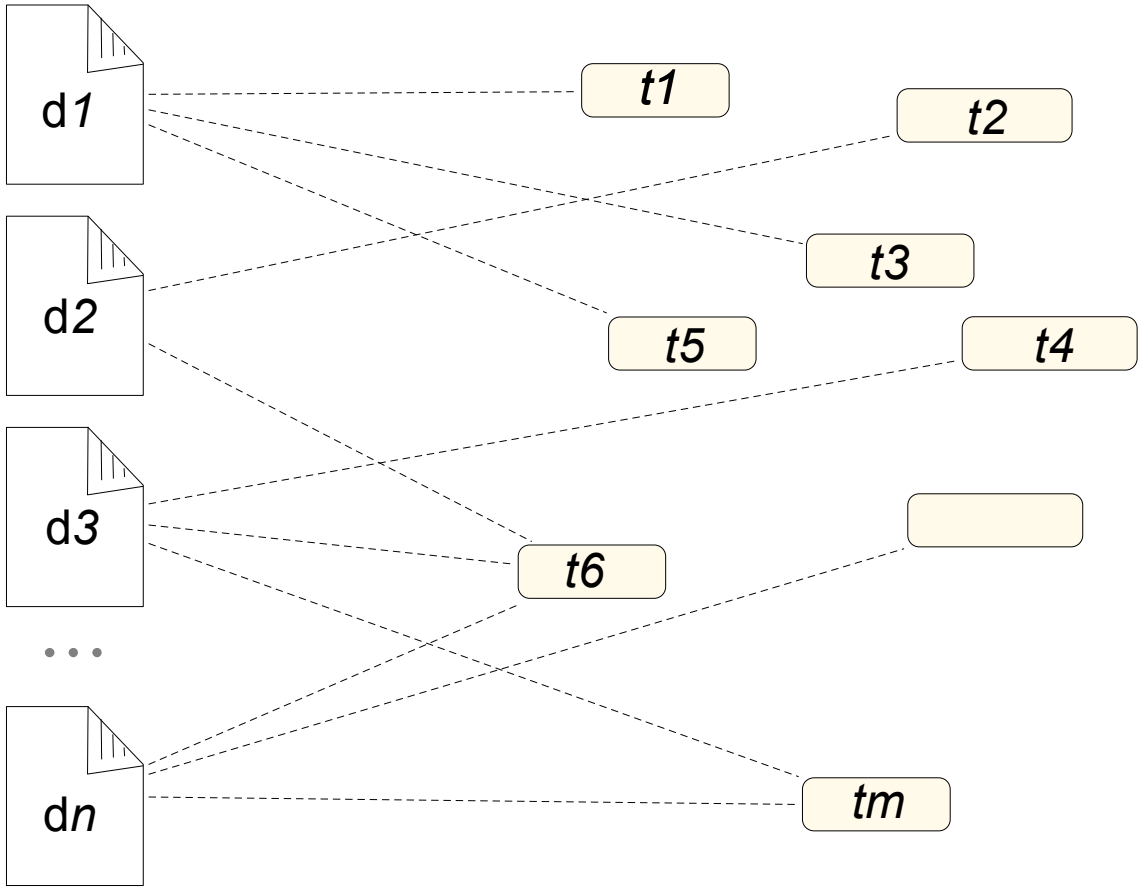


(Wong et al., 2012)

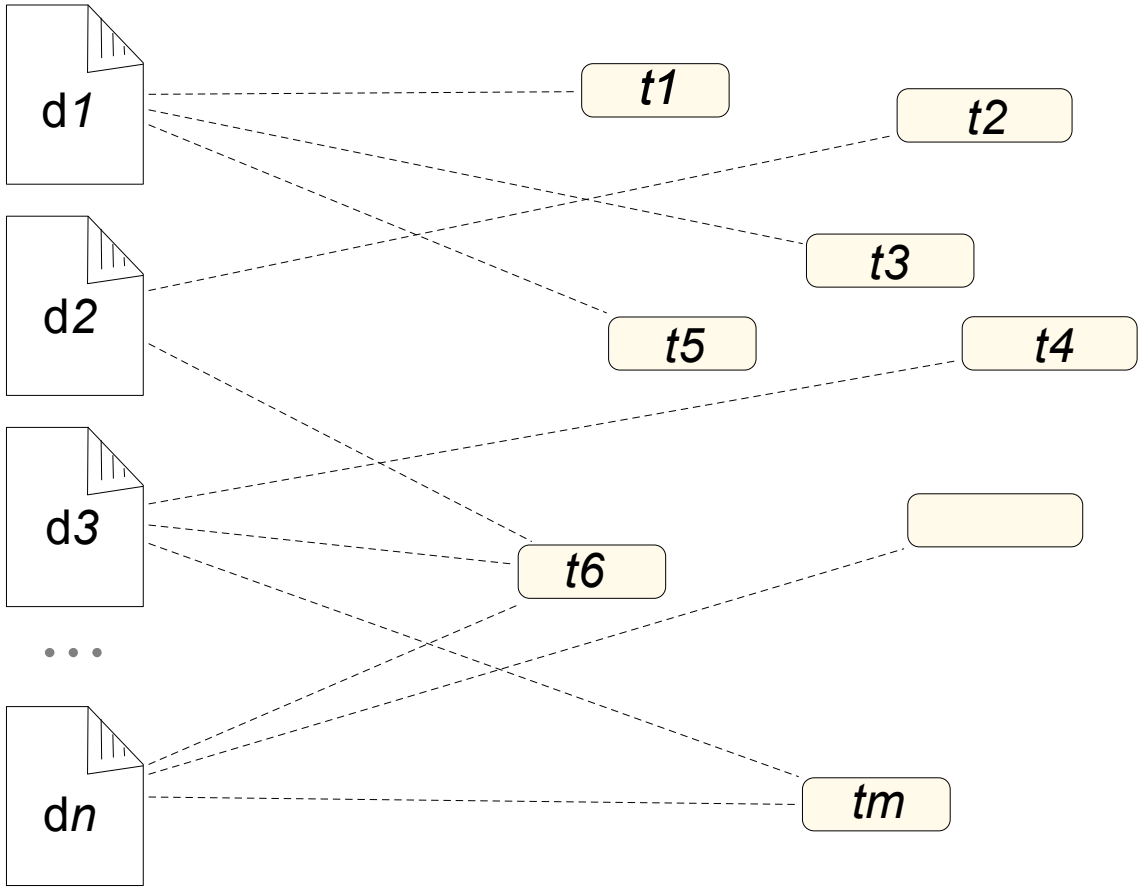
Lightweight ontológie

- Termy
  - jedno a viac-slovné výrazy
  - *pizza*
- Slovníky
  - Termy s výkladom, vysvetlením
  - *pizza* - A baked Italian dish of a thinly rolled bread dough crust typically topped before baking with tomato sauce, cheese and other ingredients such as meat, vegetables or fruit

Termy



Slovníky



t1 = a sauce made of tomatos

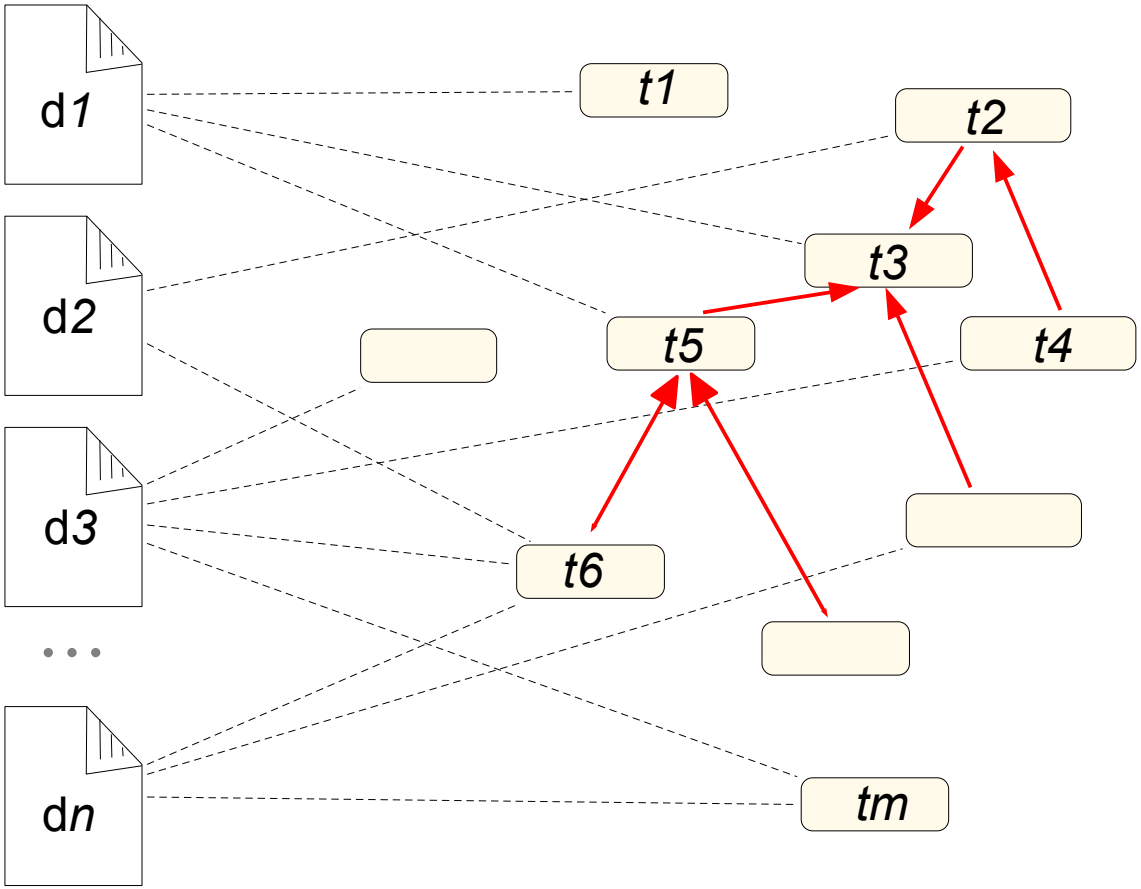
t2 = a baked italian dish …

t3 = ...

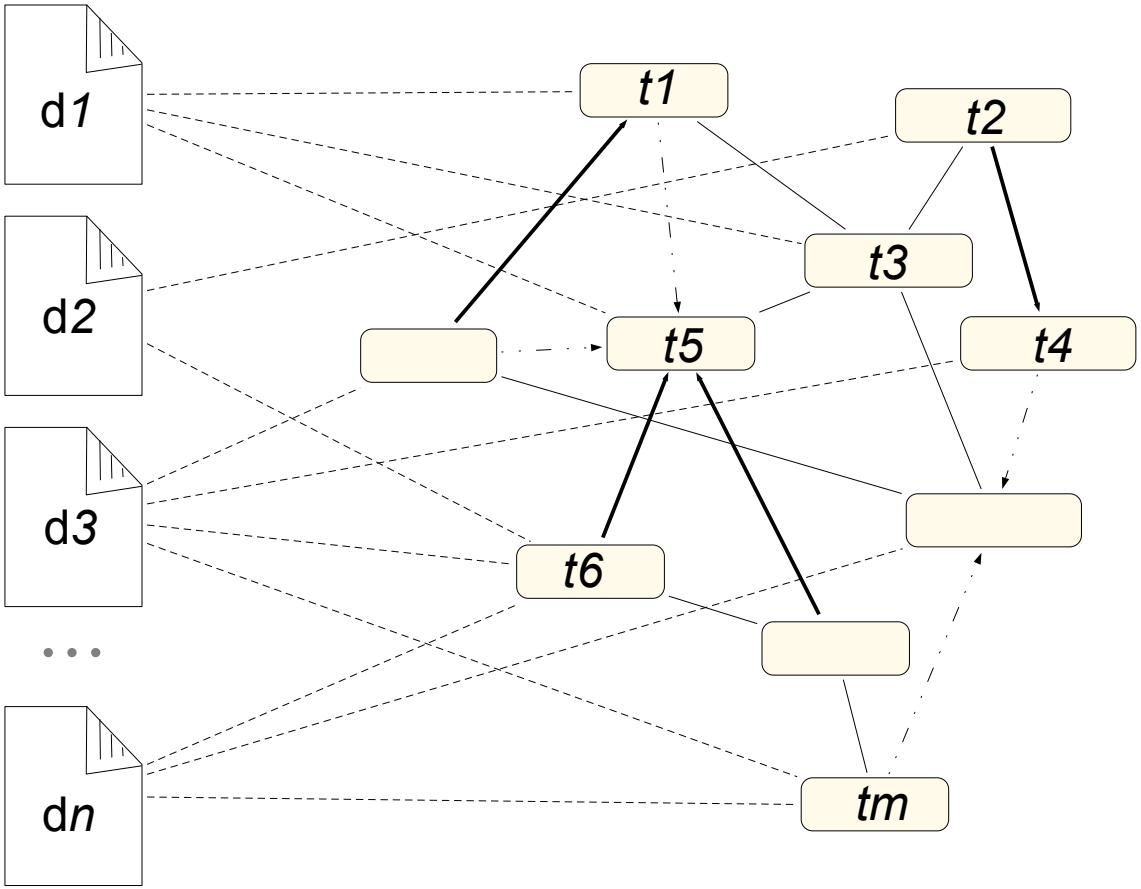
Lightweight ontológia

- Taxonómie
  - Hierarchické vzťahy typu is-a
  - *food*, *pizza*, *margherita*
- Tezaury
  - Slovníky, navyše s:
  - Hierarchickými vzťahmi: všeobecnosť (hypero-/hyponymá), časť celku (mero-/holonymá)
  - Vzťahmi ekvivalencie
  - Vzťahmi paradigmatickej podobnosti (relatedness)

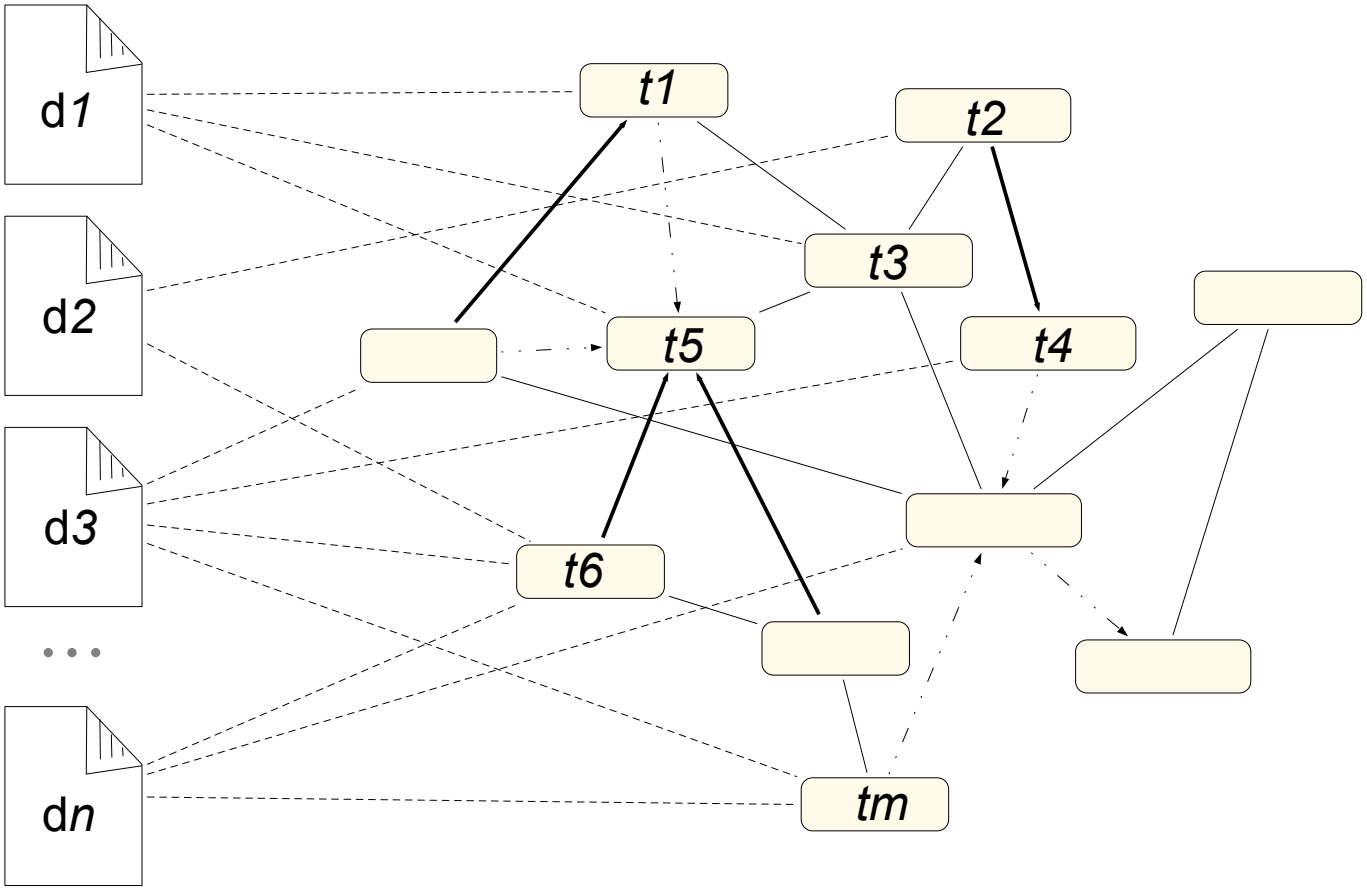
Taxonómie



Tezaury



Ľahká ontológia



Heavyweight ontológia

$$O = \{C, \leq_C, R, \sigma_R, \leq_R, A, \sigma_A, T\}$$

- four disjoint sets C, R, A and T whose elements are called concept identifiers, relation identifiers, attribute identifiers and data types, respectively,
- a semi-upper lattice $\leq_c$ on C with top element $root_c$, called concept hierarchy or taxonomy,
- function σ_R: $R \rightarrow C^+$ called relation signature,
- a partial order $\leq_R$ on R, called relation hierarchy, where $r_1 \leq_R r_2$ implies $|\sigma_R(r_1)| = |\sigma_R(r_2)|$ and $\pi_i(\sigma_R(r_1)) \leq_c \pi_j(\sigma_R(r_2))$, for each $1 \leq i \leq |\sigma_R(r_1)|$, and
- a function σ_A: $A \rightarrow C \times T$, called attribute signature,
- a set T of datatypes such as strings, integers, etc.

(Cimiano, 2006)

Heavyweight ontológia

- Definition 2 (Domain and Range)
- Definition 3 ($\mathscr{L}$-axiom System)
- Definition 4 (Lexicon)
- Definition 5 (Knowledge Base (KB))
- Definition 6 (Instance Lexicon)
- Definition 7 (Extension)
- Definition 8 (Intension)

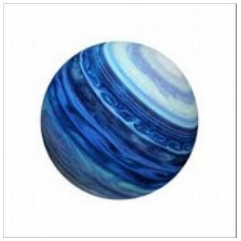
(Cimiano, 2006)

Koncept – trojica

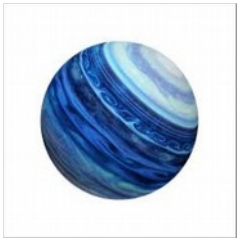
(i(c), e(c), ref(c) **)**

- i(c) – intenzionálny opis
- e(c) – extenzionálny opis
- ref(c) – lexikálna realizácia v korpuse

Príklad: Plynný obor

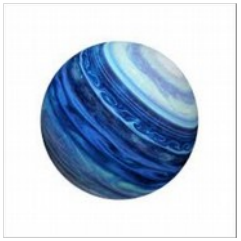


Príklad: Plynný obor



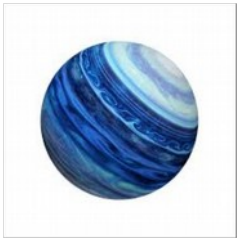
i(c):

Príklad: Plynný obor



i(c): veľká planéta, ktorá nie je zložená prevažne z hornín alebo inej pevnej látky

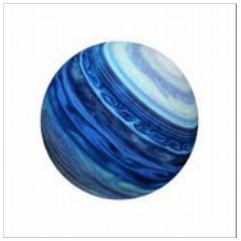
Príklad: Plynný obor



i(c): veľká planéta, ktorá nie je zložená prevažne z hornín alebo inej pevnej látky

e(c):

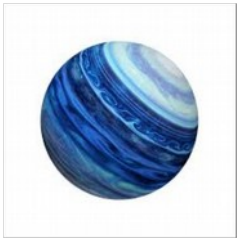
Príklad: Plynný obor



i(c): veľká planéta, ktorá nie je zložená prevažne z hornín alebo inej pevnej látky

e(c): Jupiter, Saturn, 47 Ursae Majoris c

Príklad: Plynný obor

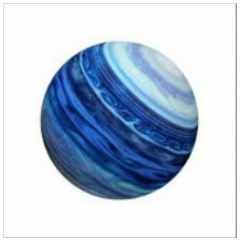


i(c): veľká planéta, ktorá nie je zložená prevažne z hornín alebo inej pevnej látky

e(c): Jupiter, Saturn, 47 Ursae Majoris c

ref(c):

Príklad: Plynný obor



i(c): veľká planéta, ktorá nie je zložená prevažne z hornín alebo inej pevnej látky

e(c): Jupiter, Saturn, 47 Ursae Majoris c

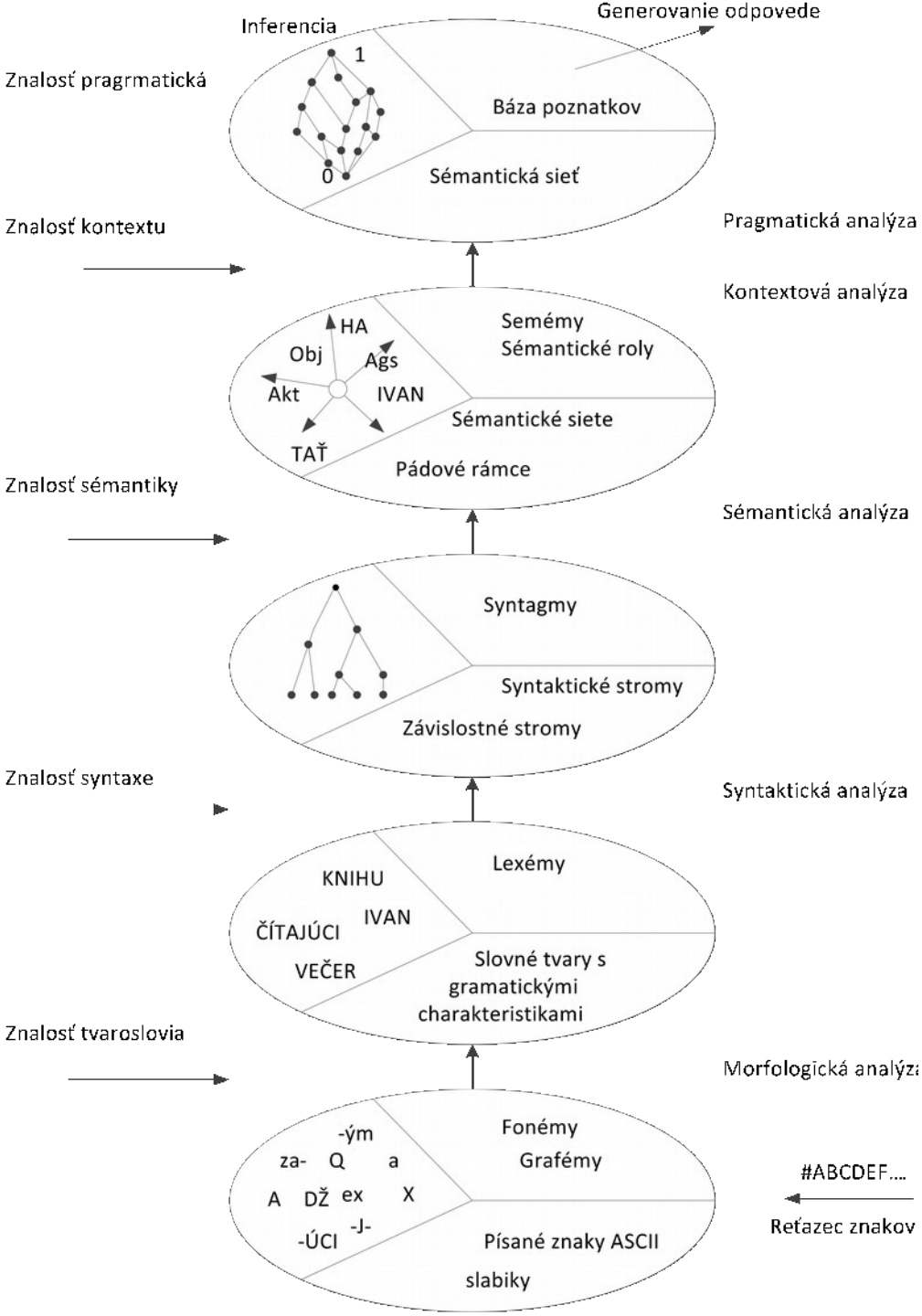
ref(c): plynný obor, Joviálna (Ioviálna) planéta

Zápis ontológií/sémantiky

- RDF
  - Resource Description Framework
- RDFS
  - RDF Schema
- OWL
  - Web Ontology Language
  - OWL Lite
  - OWL DL
  - OWL Full

Spracovanie prirodzeného jazyka

Spracovanie prirodzeného jazyka


Inferencia
Generovanie odpovede
Znalosť pragrmatická
1
0
Báza poznatkov
Sémantická sieť
Znalosť kontextu
Pragmatická analýza
Kontextová analýza
HA
Obj
Ags
Akt
IVAN
TAŤ
Semémy
Sémantické roly
Sémantické siete
Pádové rámce
Znalosť sémantiky
Sémantická analýza
Syntagmy
Syntaktické stromy
Závislostné stromy
Znalosť syntaxe
Syntaktická analýza
KNIHU
IVAN
ČÍTAJÚCI
VEČER
Lexémy
Slovné tvary s gramatickými charakteristikami
Znalosť tvaroslovia
Morfologická analýza
-ým
za-
Q
a
A
DŽ
ex
X
-ÚCI
-J-
Fonémy
Grafémy
Písané znaky ASCII
slabiky
#ABCDEF....
Reťazec znakov

- Zelené žaby žijú v rybníku.

- Červené žaby majú dlhé nosy.

- Zelené žaby žijú v rybníku.

- Červené žaby majú dlhé nosy. # pragmaticky
- Červené idey majú dlhé nosy.

- Zelené žaby žijú v rybníku.

- Červené žaby majú dlhé nosy. # pragmaticky
- Červené idey majú dlhé nosy. # sémanticky
- Červenými ideou mať nos dlhý.

- Zelené žaby žijú v rybníku.

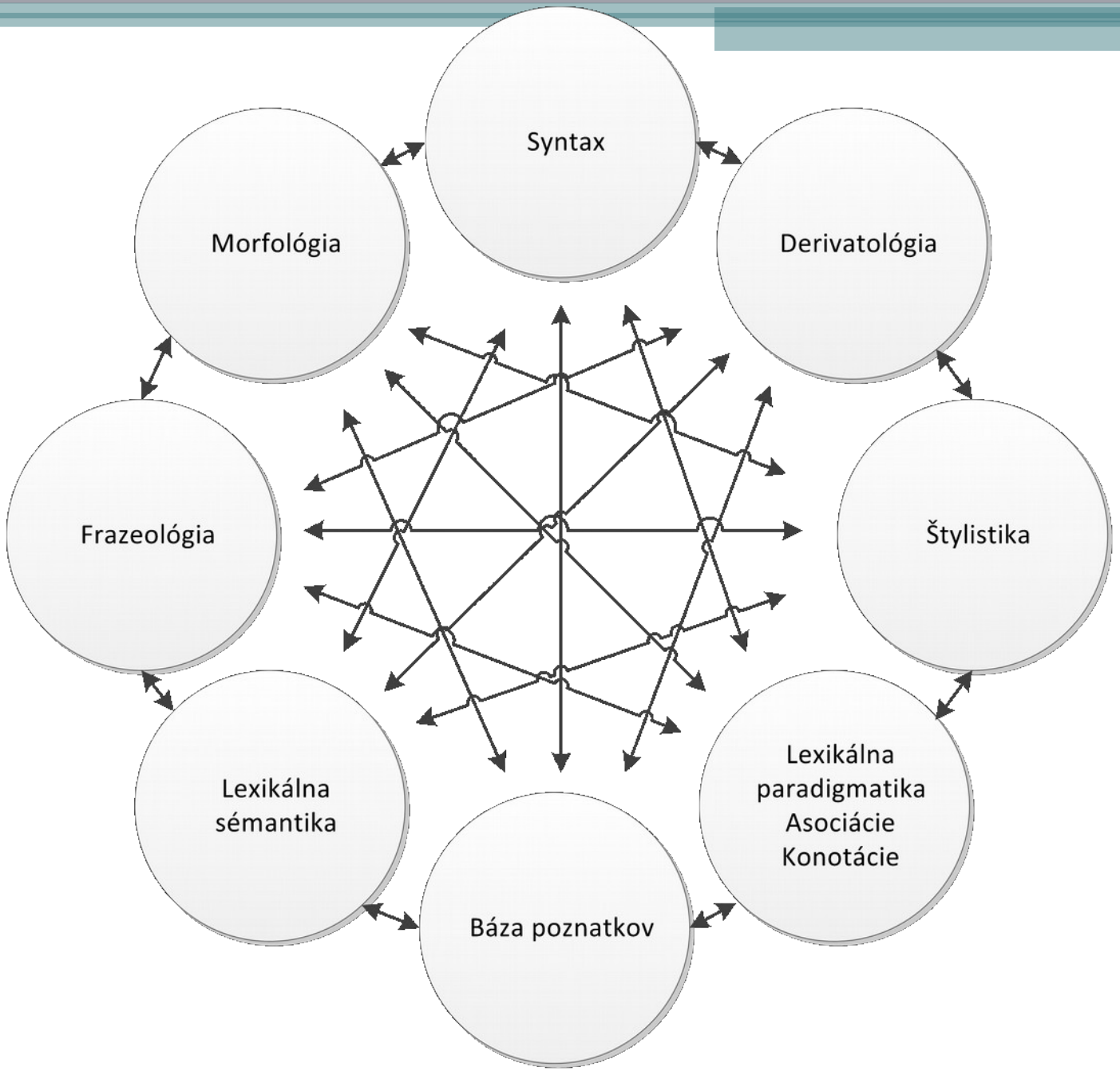
- Červené žaby majú dlhé nosy. # pragmaticky
- Červené idey majú dlhé nosy. # sémanticky
- Červenými ideou mať nos dlhý. # syntaktický
- Červenskami ideová maty nosník prídlžie.

- Zelené žaby žijú v rybníku.

- Červené žaby majú dlhé nosy. # pragmaticky
- Červené idey majú dlhé nosy. # sémanticky
- Červenými ideou mať nos dlhý. # syntaktický
- Červenskami ideová maty nosník prídlžie. # morfematicky
- Cfvskmn idée math noss ďlha.

- Zelené žaby žijú v rybníku.

- Červené žaby majú dlhé nosy. # pragmaticky
- Červené idey majú dlhé nosy. # sémanticky
- Červenými ideou mať nos dlhý. # syntaktický
- Červenskami ideová maty nosník prídlžie. # morfematicky
- Cfvskmn idée math noss ďlha. # fonologicky


Syntax
Morfológia
Derivatológia
Frazeológia
Štylistika
Lexikálna sémantika
Lexikálna paradigmatika
Asociácie
Konotácie
Báza poznatkov

Spracovanie prirodzeného jazyka

- Predspracovanie textu
- Určovanie slovných druhov
- Lematizácia a stemovanie
- Rozpoznávanie vlastných pomenovaní
- Identifikácia viacslovných pojmov
- Rozpoznanie koreferencií
- Syntaktická analýza

Predspracovanie textu

1. Konverzia dokumentov na jednotné kódovanie
2. Extrakcia čistého textu
3. Tokenizácia a segmentácia
4. Normalizácia
5. Odstránenie/označenie stop slov

text.fiit.stuba.sk

Predspracovanie textu: Tokenizácia

Vstup: Joviálna (Ioviálna) planéta alebo plynný obor je planéta, ktorá je svojou veľkosťou a zložením podobná Jupiteru.

Výstup: [Joviálna][ ][(][Ioviálna][)][ ][planéta][ ] [alebo][ ][plynný][ ][obor][ ][je] [ ] [planéta][,][ ][ktorá][ ][je][ ][svojou][ ] [veľkosťou][ ][a][ ][zložením][ ] [podobná][ ][Jupiteru][.]

Určovanie slovných druhov

Vstup: 'rybníkom'

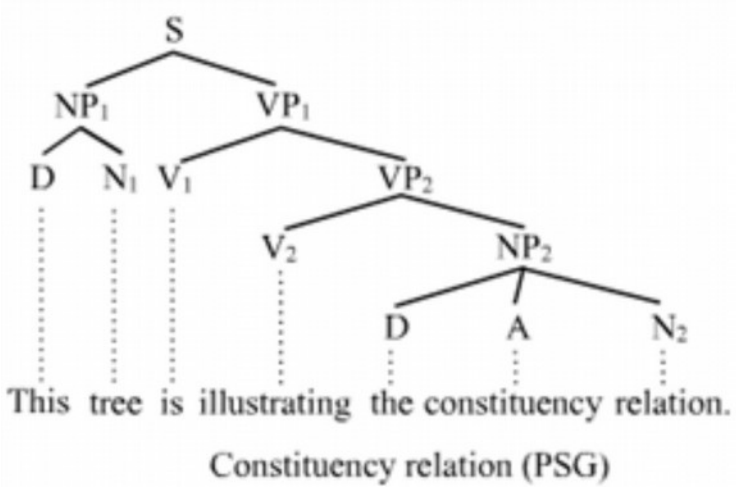
Výstup: 'podstatné meno, mužský rod, jednotné číslo, inštrumentál' (a niekedy aj viac)

Rozpoznanie koreferencií

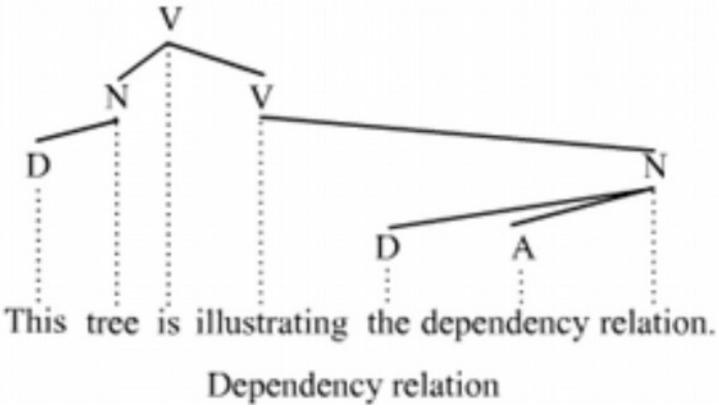
Monika miluje tenis. Dokáže obetovať hodiny, aby ho trénovala.

Rytier zosadol z koňa, aby sa napil.

Syntaktická analýza



Constituency relation (PSG)



Dependency relation