PLIN021 SÉMANTICKÁ ANALÝZA V PRAXI

ZUZANA NEVĚŘILOVÁ

2020/21

LEXIKÁLNÍ ZDROJE PRO ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA

DEFINICE VÝZNAMU

VÝZNAM A UŽITÍ
V KONTEXTU

ONTOLOGIE

„formální a explicitní specifikace sdílené konceptualizace“

(Gruber, 2009)

- Formální = formálně (matematicky) popsatelná
- Explicitní = nic není zamlčeno
- Specifikace = popis
- Sdílená = lidé by se na ní shodli
- Konceptualizace = tvorba pojmů a jejich vyjádření v jazyce

ONTOLOGIE

„formální a explicitní specifikace sdílené konceptualizace“

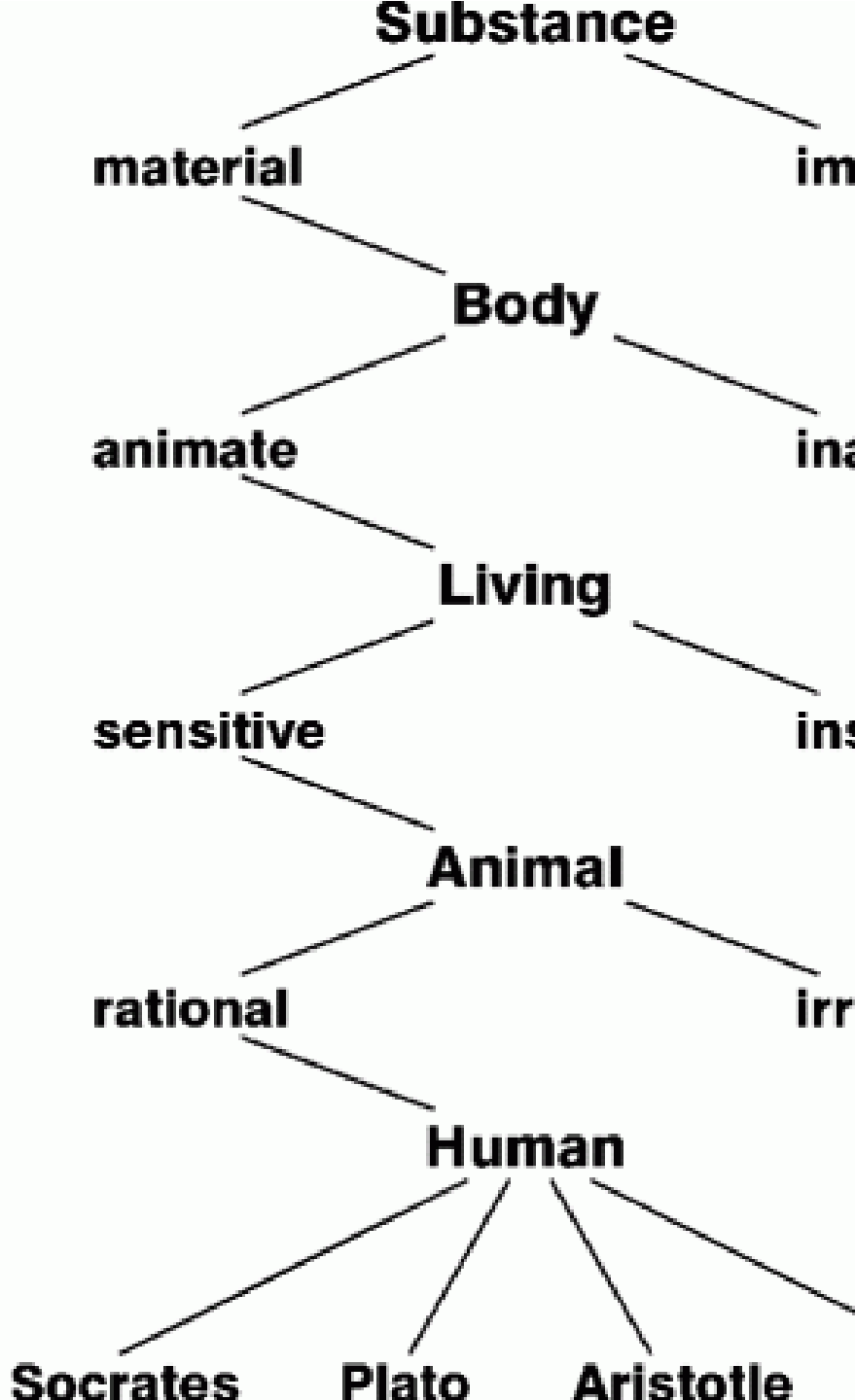
(Gruber, 2009)

- Inventář významů (slovník, glosář, rejstřík, seznam)
- Inventář relací (taxonomie)

TAXONOMIE

- Aristoteles -- kategorie (všech) entit, které mohou lidé vnímat
- Porfyrios -- uspořádal kategorie
- Carl Linné -- klasifikace (všech) organismů

Důležité rysy: **uzly** jsou třídy (organismů, entit . . .), **třídy** jsou strukturované do **stromu** (podtřída, nadtřída), uzly na stejné úrovni se vzájemně vylučují (implicitní předpoklad)

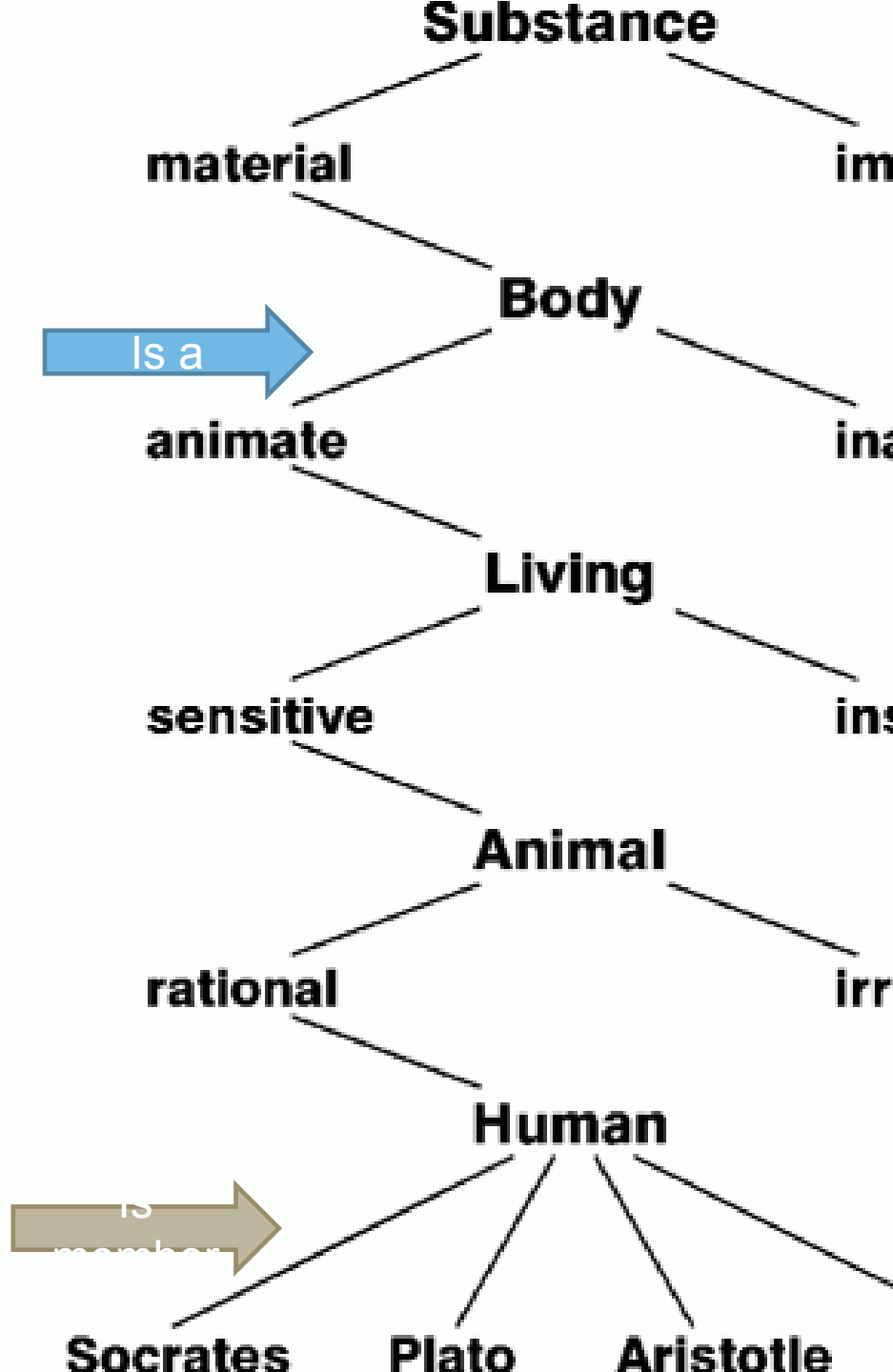


TAXONOMIE

- Třída – generický popis skupiny jednotlivců
- Instance – jednotlivec

Pes je masožravec, nepohrdne však ani ovocem.
Alík má rád švestky.

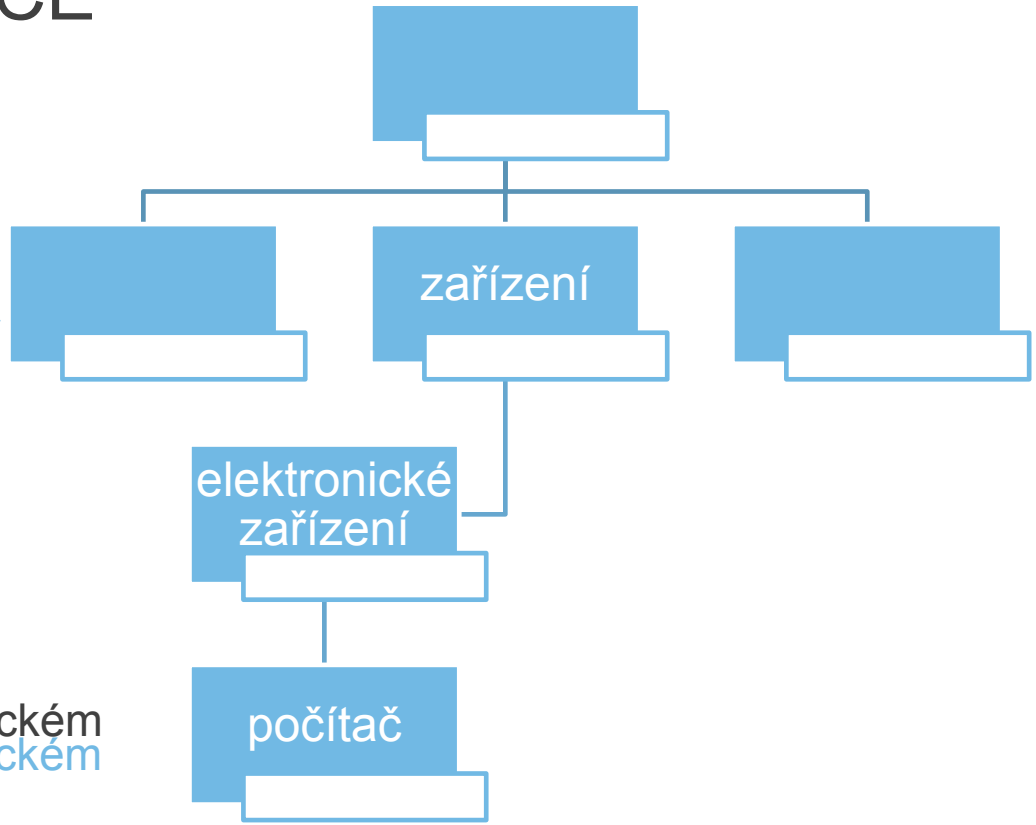
Americký prezident je zároveň předsedou vlády.
Americký prezident má babičku v Africe.



TAXONOMIE A SLOVNÍKOVÉ DEFINICE

klasická definice =

genus proximum + differentia specifica

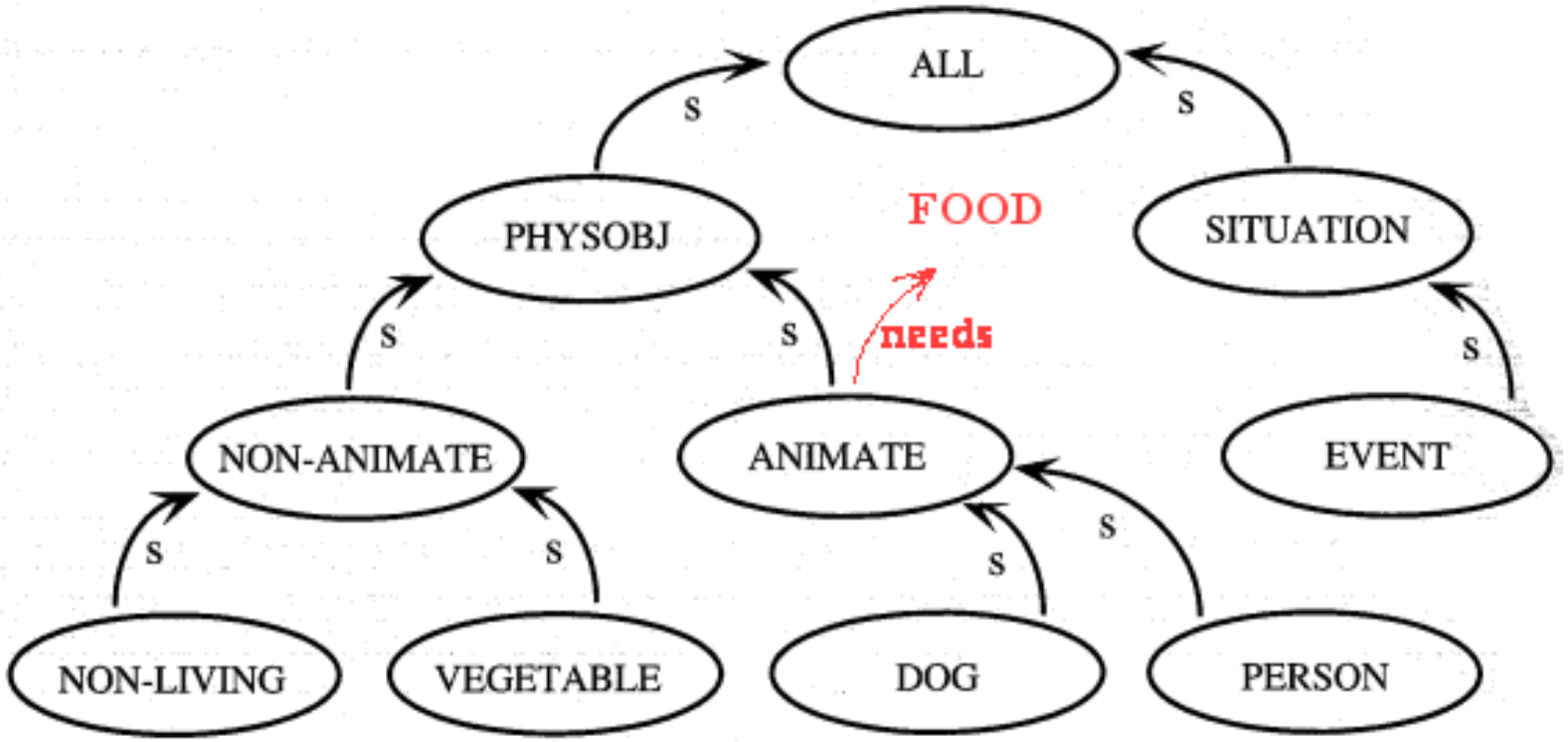


- Počítač je v informatice elektronické zařízení, které zpracovává data pomocí předem vytvořeného programu.
- Elektronické zařízení je zařízení, jehož funkce závisí na elektrickém proudu nebo na elektromagnetickém poli.

OD STROMU K SÍTI

- Více typů relací
- Odvozování

X platí pro třídu → X platí



SÉMANTICKÁ SÍŤ A RELACE

- nadtyp–podtyp, is a, is-a, isa (hypo/hyperonymie)
- instance třídy, member of
- část–celek, has a (holo/meronymie)
- upřesnění akce (troponymie)
- opozitnost (antonymie)
- příčina–následek
- ...

ODVOZOVÁNÍ VE ZNALOSTNÍCH BÁZÍCH

- Fakt *F* = tvrzení s pravdivostní hodnotou (např. ptáci létají)
  Báze znalostí (knowledge base) *KB* = (pokud možno konzistentní) soubor faktů (např. ptáci létají, vlaštovka je pták)

- Pokud z *KB* plyne *F* a přidáme další fakt takový, že KB je stále konzistentní, je *KB* **monotónní** reprezentace. (Allen, 1995)

ptáci létají vlaštovka je pták	ptáci létají tučňák je pták	ptáci kromě tučňáků létají tučňák je pták
vlaštovka létá	tučňák létá	

ODVOZOVÁNÍ VE ZNALOSTNÍCH BÁZÍCH

ptáci kromě tučňáků, mláďat, pštrosů, kiwi, mrtvých ptáků atd. létají
X je pták (kromě tučňáků, mláďat, pštrosů, kiwi, mrtvých ptáků atd. létají)

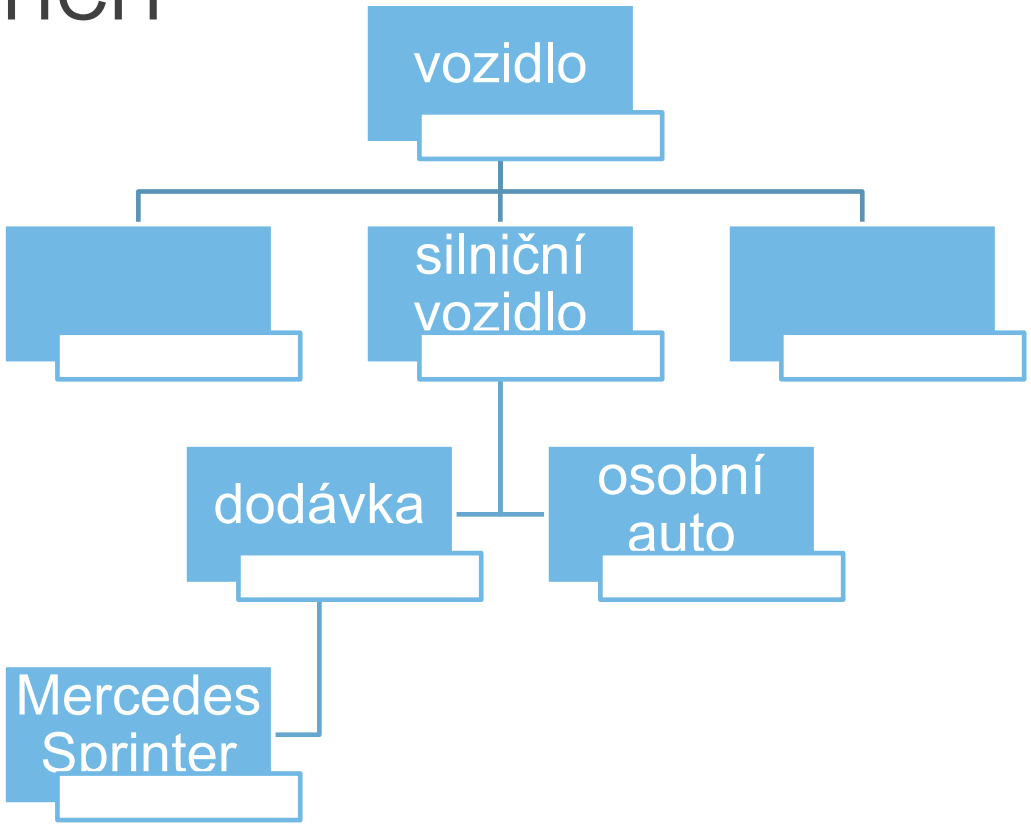
X létá

- V přirozeném jazyce potřebujeme **nemonotónní odvozování**.
- Použijeme tzv. **implicitní pravidlo** (default rule), které platí vždy, když neexistuje sporné pravidlo pro podtřídu.

Ptáci létají. Tučňák je pták. Tučňák nelétá.

ODVOZOVÁNÍ V SÉMANTICKÝCH SÍTÍCH

- silniční vozidlo má (has part) volant
- dodávka je (isa) silniční vozidlo
- dodávka má (has part) volant
- Mercedes Sprinter je (member of) dodávka
- Mercedes Sprinter má (has part) volant



EXISTUJÍCÍ SÉMANTICKÉ SÍTĚ

- WordNet (lexikální síť)
  - Původně Princeton WordNet (PWN) pro angličtinu: 117 tisíc synsetů
  - EuroWordNet – evropské jazyky, vrcholová ontologie (Top Ontology), základní koncepty (Base Concepts), interlingual identifier (ILI)
  - BalkaNet – balkánské a evropské jazyky
  - Global WordNet Association – sdružuje WordNety pro všechny jazyky
  - Český WordNet (FI MUNI) – 28 tisic synsetů, automaticky rozšířená verze: 40 tisíc synsetů
  - Největší WordNet – polská Słowosieć: 294 tisíc synsetů

EXISTUJÍCÍ SÉMANTICKÉ SÍTĚ

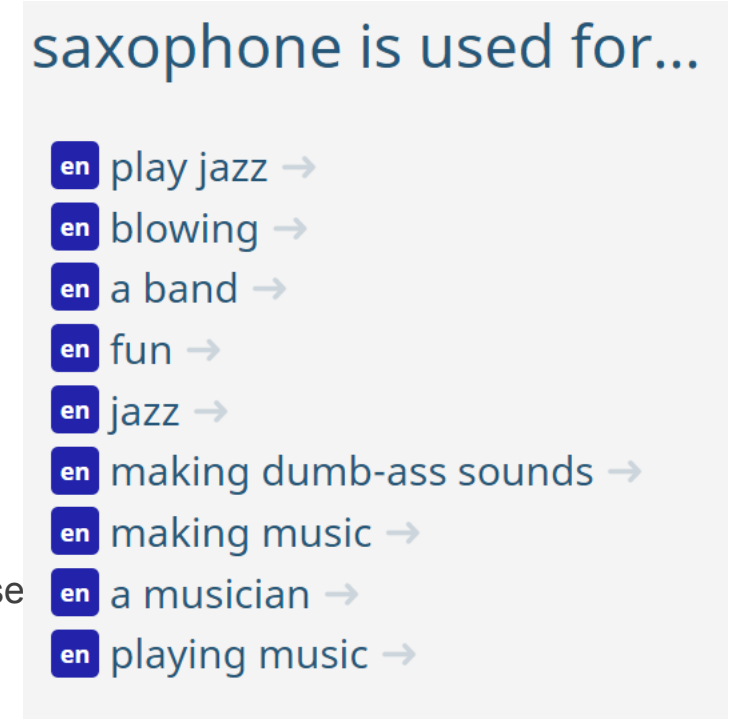
- Wikipedia a hlavně její strojově zpracovatelná varianta dbPedia
  - (Polo)automaticky vygenerovaná znalostní báze
  - Základem jsou tzv. Infoboxy z Wikipedie, kategorizace článků, obrázky, geografické souřadnice a externí odkazy
  - Výhodou je, pokud wikipedista používá šablony stránek
  - Poskytuje rozhraní přes Virtuoso SPARQL, kde je možné zadávat komplexní dotazy
  - Obsahuje data z více než 100 Wikipedií
  - Propojení s jinými ontologiemi

Římskokatolická farnost Těšetice u Olomouce

farní kostel

	Základní údaje
Děkanát	Olomouc
Diecéze	arcidiecéze olomoucká
Provincie	moravská
Farář	R. D. Mgr. Mirosław Łukasiewicz
	Území farnosti

EXISTUJÍCÍ SÉMANTICKÉ SÍTĚ



- ConceptNet – common sense knowledge
  - Data vznikla pomocí crowdsourcingu v projektu Open Mind Common Sense
  - Motivace projektu byla chybějící informace pro systémy s umělou inteligencí – common se
  - Obsahuje tvrzení o běžných věcech ve formě trojic, které připomínají věty
  - 300 000 uzlů, 1,6 milionu hran (2004)
  - Po propojení všech dostupných jazyků (přes Wiktionary): 8 milionů uzlů, 21 milionů hran, 83 jazyků s alespoň 10 000 hranami

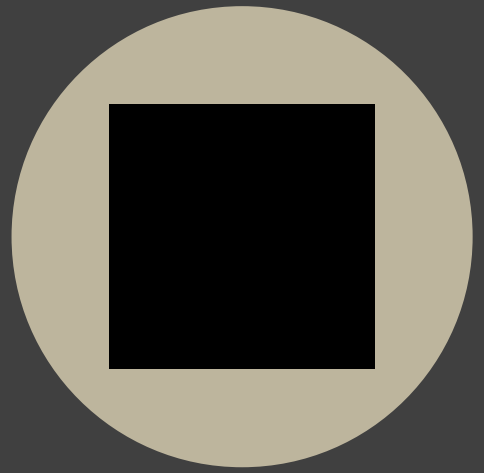
EXISTUJÍCÍ SÉMANTICKÉ SÍTĚ

- SUMO/MILO – Suggested Upper Merged Ontology/Mid-Level Ontology
- YAGO – Yet Another Great Ontology
- NELL – Never-Ending Language Learning
- Doménové ontologie: GeoNames, Wikidata, Global Research Identifier Database, ...

Linked (Open) Data

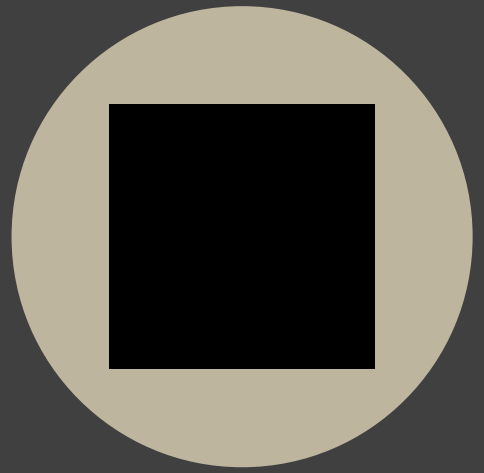
LEXIKÁLNÍ ZDROJE PRO ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA

DEFINICE VÝZNAMU



VÝZNAM A UŽITÍ
V KONTEXTU

ZDROJE ZNALOSTÍ PRO ZPRACOVÁNÍ PŘIROZENÉHO JAZYKA



ZNALOST SVĚTA

ZNALOST JAZYKA

LITERATURA

- Gruber, Thomas. **Toward Principles for the Design of Ontologies Used for Knowledge Sharing**. *International Journal Human-Computer Studies* Vol. 43, Issues 5-6, Novemer 1995, p.907-928. https://tomgruber.org/writing/onto-design.pdf
- Gruber, Thomas. **Ontology**. In Liu, L. and Özsu, M. T., editors, *Encyclopedia of Database Systems*, Springer Verlag. 2009. pp 1963–1965. https://link-springer-com-443.webvpn.zisu.edu.cn/referenceworkentry/10.1007/978-1-4614-8265-9_1318
- Liu, H., Singh, P. **ConceptNet — A Practical Commonsense Reasoning Tool-Kit**. *BT Technology Journal* **22,** 211–226 (2004). https://doi.org/10.1023/B:BTTJ.0000047600.45421.6d http://alumni.media.mit.edu/~hugo/publications/papers/BTTJ-ConceptNet.pdf
- Fellbaum, Christiane: *WordNet: An Electronic Lexical Database*. Bradford Books. 1998.
- Allen, J. *Natural Language Understanding (2nd ed.)*. Benjamin-Cummings Publishing Co., Inc., Redwood City, CA, USA. 1995