Modelování ontologií (Ontology Modeling)

...

ISKM89 Organizace dat - sémantický web | podzim 2023

Zuzana Nevěřilová | Centrum zpracování přirozeného jazyka

Modelování

- Model je analogií (části) reálného světa. (B. Russell)
- Sémantika modelu - formální interpretace modelu (A. Tarski)

Interpretace: relace mezi syntaxí a sémantikou

A: $\forall x{:}P(f(x), x)$

B: $\exists x{:}P(f(x), x)$

C: $P(f(x), x)$

Interpretace I: $U=\mathbb{N}$, $f \rightarrow x^2$, $P \rightarrow$ relace >

Formule A v I není pravdivá.
Formule B v I je pravdivá.
Formule C je v I splnitelná, ale není pravdivá.

Proč vytvářet ontologie?

To **share** common understanding of the structure of information among people or software agents

To enable **reuse** of domain knowledge

To make domain assumptions **explicit**

To separate **domain** knowledge from the operational knowledge

To **analyze** domain knowledge

https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

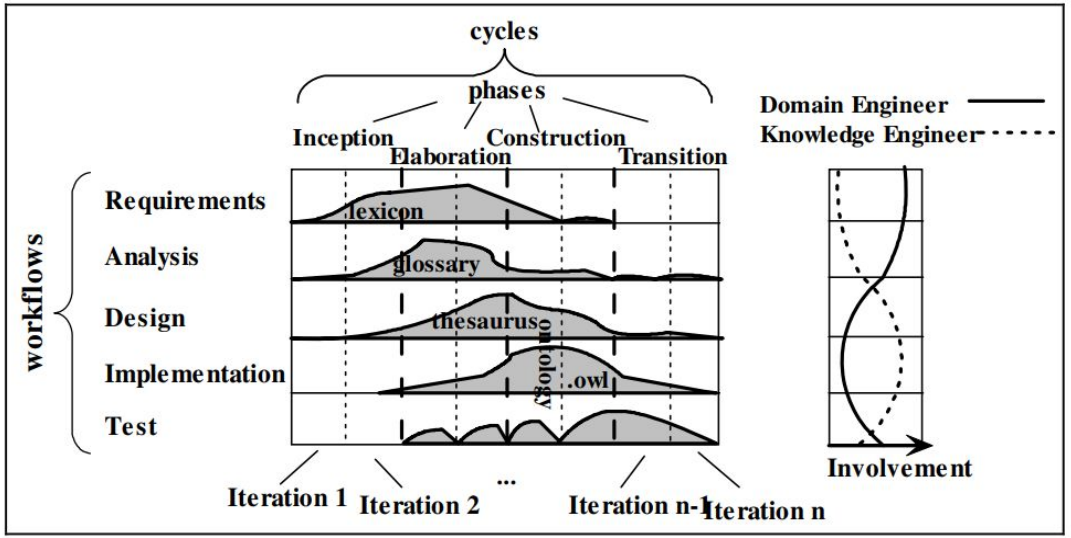
Jak vytvářet ontologie?

Metodologie Ontology Development 101

určit rámec	znát existující	vyjmenovat termíny	definovat třídy	definovat vlastnosti	definovat omezení	vytvořit instance
Na jaké otázky ontologie odpovídá? Jaké situace ontologie popisuje?	Může to být úsporné. Nástroje pro existující ontologie už jsou hotové.	Jaké jsou pojmy v doméně? Jaké vlastnosti mají naše pojmy? Co o nich chceme v ontologii uchovávat?	Jaké vlastnosti jsou společné pro různé objekty? Lze uspořádat třídy do hierarchie?	Které vlastnosti budou objektové, tj. relace k jinému objektu? Které budou datové, tj. literály?	Datové vlastnosti mohou mít jen určité typy. Objektové vlastnosti mají doménu, obor hodnot a možná i určitou kardinalitu.	Lze instance vyhledat? Popisují instance to, co jsme na začátku chtěli mít popsané?

https://protege.stanford.edu/publications/ontology_development/ontology101.pdf

Unifikovaný proces budování ontologií (Unified Process of Ontology Building)

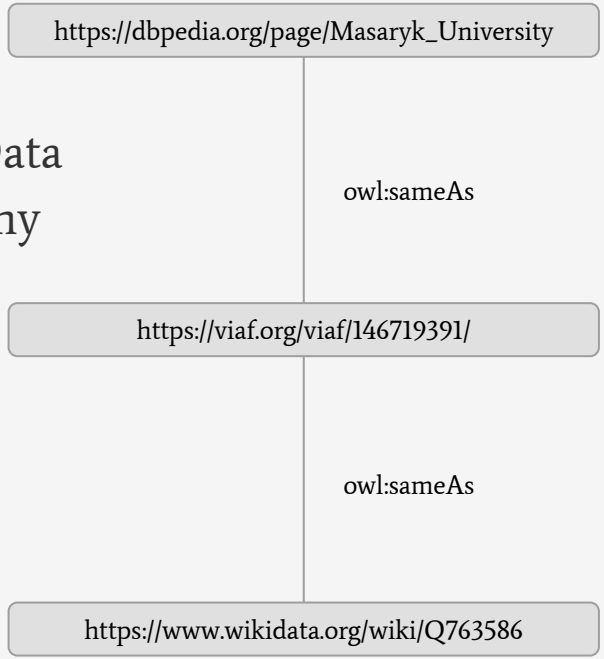


https://link.springer.com/chapter/10.1007/11546924_64

Jak začít?

- zjistit, jaké domény potřebujeme popsat
- zjistit, jaké ontologie pro domény existují
- zvolit vhodná URI - čtyři pravidla pro Linked Data
- vytvořit slovníky (vocabulary) pro důležité pojmy
  - třídy, atributy, relace, instance, pravidla, omezení
- vygenerovat odkazy
- popsat metadata (DublinCore, Semantic Web Publishing Vocabulary)
- použít vhodné publikační nástroje

http://wbsg.informatik.uni-mannheim.de/bizer/wiqa/swp/SWP-UserManual.pdf



Otevřená propojená data (Linked Open Data)

Tim Berners-Lee (2006): Čtyři pravidla pro otevřená propojená data

- použij URI pro jména věcí
- použij URI se schématem HTTP, aby lidé věci našli na Internetu
- když někdo požaduje URI, přidej užitečné informace (s využitím standardů RDF a SPARQL)
- přidej odkazy na další URI, ať mohou ostatní objevovat další věci

https://www.w3.org/DesignIssues/LinkedData.html

Co je důležité vědět

třídy, **atributy** tříd, **relace**, **instance**, **pravidla**, **omezení** (constraints) - co bude co

- modelování je subjektivní proces
- není jeden správný způsob, jak modelovat doménu
- je vhodné postupovat od aplikace (i návrhu) a možných budoucích rozšíření
- vývoj ontologie je iterativní proces
- koncepty v ontologii by měly být pojmenované tak, jak se objekty (fyzické či logické) v doméně pojmenovávají - podstatná jména pro objekty, slovesa pro vlastnosti a vztahy

Modelování ontologie od začátku

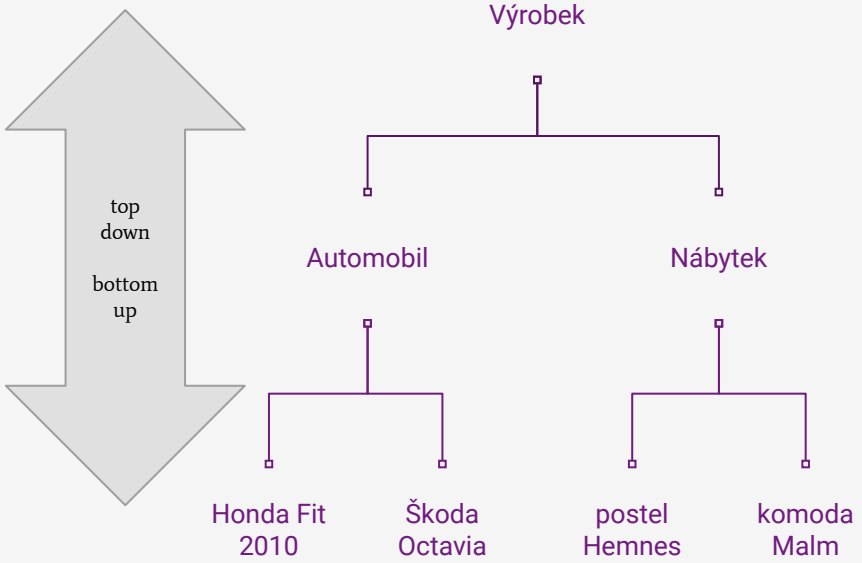
Doména:

- jaké domény má ontologie pokrýt?
- na jaké otázky by ontologie měla poskytovat odpovědi?
- kdo bude ontologii používat a kdo ji bude spravovat?
- lze použít (importovat) existující ontologie?

Modelování ontologie od začátku

Slovník

- jaké jsou důležité pojmy v doméně?
  - jednoduchý výpis
  - rozdělení na třídy a vlastnosti (properties)
  - taxonomie (hierarchie) tříd i vlastností
  - objektové vlastnosti (object properties)
    definice pomocí slotů
    židle - sedák - sedět - pohodlí - opírat
  - upřesnění slotů (facets)
    doména a obor hodnot (domain, range)
    kardinalita
    počet nohou, počet sedáků, sedět (kdo?)
- jaké jsou instance v doméně?



Modelování ontologie z databáze

Entity Relationship Model (ERM)
entitně vztahový model
(přístup bottom up)

výsledkem modelu je ER diagram

softwarové nástroje pro modelování:

- MS Visio (spec. licence)
- draw.io (https://drawio-app.com/blog/entity-relationship-diagrams-with-draw-io/)
- smartdraw (https://www.smartdraw.com/entity-relationship-diagram/er-diagram-tool.htm)
- ...

z ER diagramu lze vygenerovat OWL schéma

https://cs.wikipedia.org/wiki/Entity-relationship_model

DOI 10.1109/ICSC.2009.61: https://ieeexplore.ieee.org/document/5298643

Vygenerování odkazů

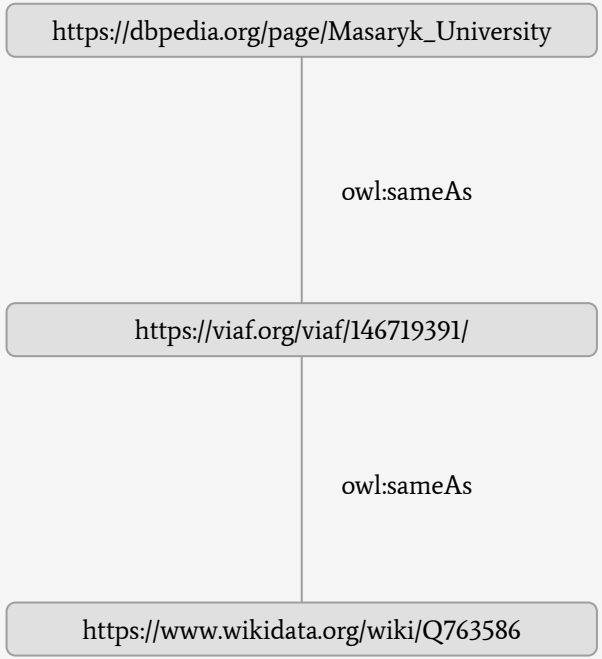
některá přiřazení je možné udělat ručně

pro automatické metody je dobré využít některou z velkých ontologií přes API:

- DBPedia
- GeoNames
- Wikidata

Předtím musí být data vyčištěná.

OpenRefine zvládne obojí.



Modelování a logika

Jaká pravidla platí globálně pro vlastnosti a třídy?

- hierarchie (podvlastnosti)
- reflexivita, symetrie, tranzitivita
- minimální a maximální kardinalita
- funkcionální vlastnost (kardinalita 1)
- existuje inverzní vlastnost?
- rozdílnost (disjointWith)
- hierarchie (podtřídy)

Co je užitečné si rozmyslet

OWL nikde neodděluje A-Box a T-Box. U některých konstruktů je oddělení složité.

Nicméně, definice tříd je spíš T-Box, individua jsou spíš A-Box.

```
ex:c rdfs:subClassOf ex:d .
```

- může být relace mezi třídami
- může být vlastnost individuí

Co je užitečné si rozmyslet

Oddělit A-Box a T-Box?

https://www.mkbergman.com/489/ontology-best-practices-for-data-driven-applications-part-2/

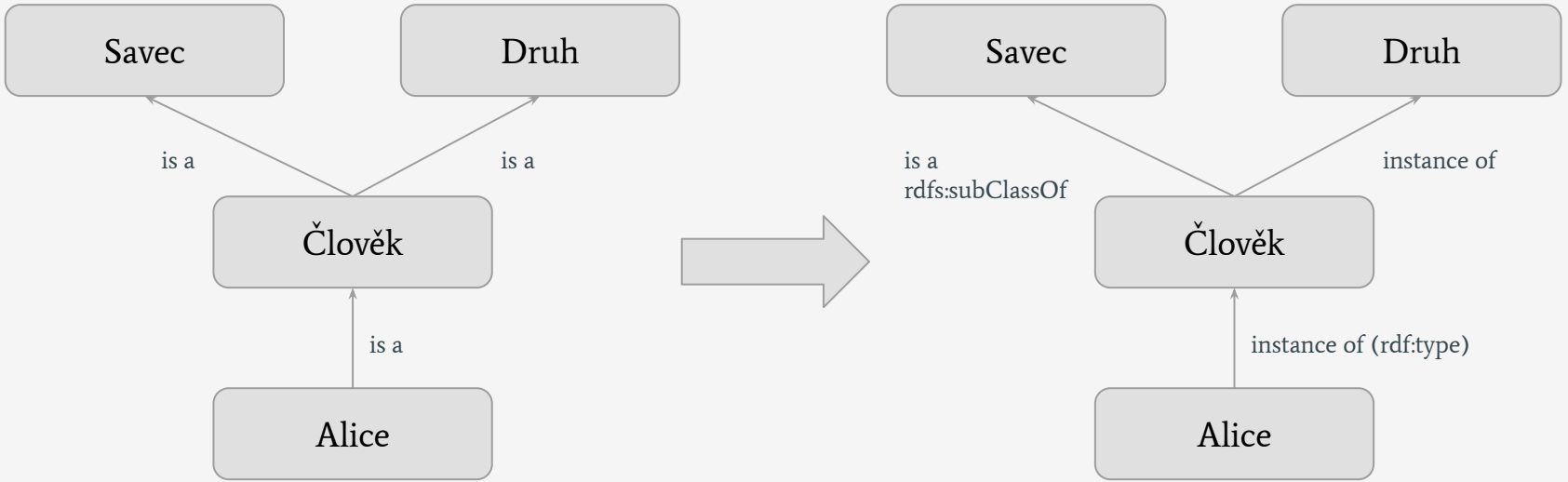
Nicméně, definice tříd je spíš T-Box, individua jsou spíš A-Box.

ex:2010_Honda_Fit_Base_Model owl:instanceOf ex:2010_Honda_Fit

ex:4B2_9999 owl:instanceOf ex:2010_Honda_Fit_Base_Model

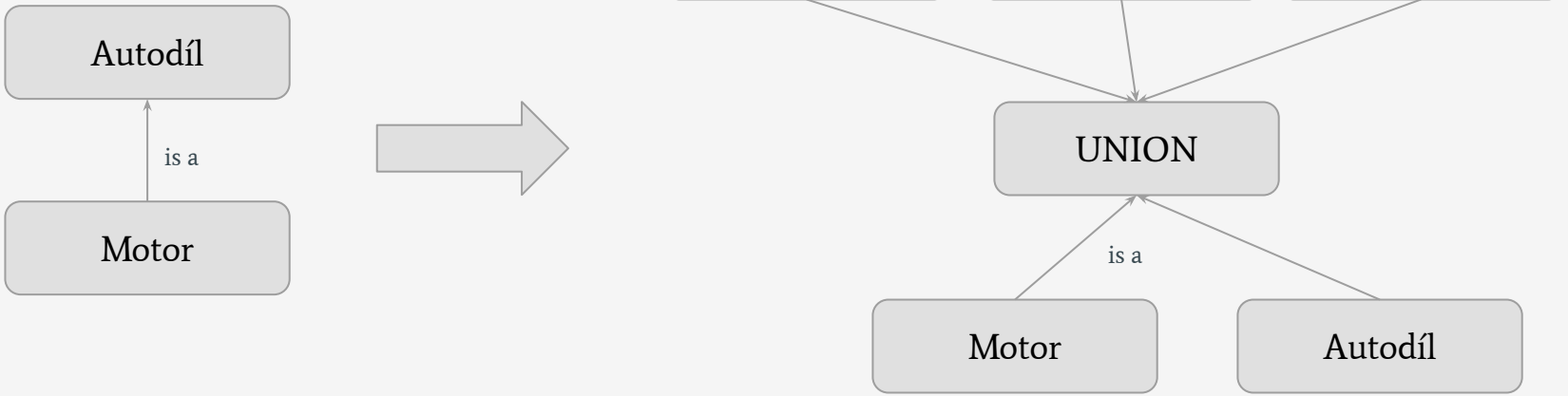
Co je užitečné si rozmyslet - a jak se vyhnout chybám

Instance a podtřídy



Co je užitečné si rozmyslet - a jak se vyhnout chybám

Podtřídy a disjunkce



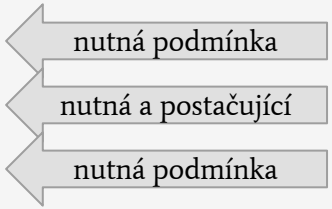
Co je užitečné si rozmyslet - a jak se vyhnout chybám

Podtřídy a konstituenty

Co je užitečné si rozmyslet - pravidla (constraints)

Je výhodné specifikovat třídy a vlastnosti co neúplněji.

```
<owl:Class rdf:ID="Human">
  <rdfs:subClassOf rdf:ID="Animal"/>
  <owl:equivalentClass rdf:ID="SmartChimp"/>
  <owl:disjointWith rdf:ID="Chimp"/>
</owl:Class>
```



- omezení v hierarchii
- omezení vlastností
- omezení kardinality

https://www.w3.org/2001/sw/BestPractices/

https://www.w3.org/2001/sw/BestPractices/OEP/ - Ontology Engineering and Patterns Task Force (OEP)

Anonymní uzly? Anonymní třídy?

Anonymní uzly obecně je lepší nemít.

Anonymní třídy jsou v pořádku při popisu ontologie (A-Box).

Anonymní třídy jsou super tam, kde není třeba komplexní koncept pojmenovávat (příliš mnoho jmen by hrozilo víceznačností nebo zmatkem).

Příklad: geografický bod (šířka, délka)

Principy modelování ontologií - co je dobré dodržet

- Strukturální interoperabilita - RDF umožňuje práci s tzv. strukturovanými i nestrukturovanými daty. Ideální je zobrazovat HTML ve webovém prohlížeči a RDF pro RDF klient.
- Federace - dotazy mohou jít přes několik provázaných datových zdrojů
- Konceptuální interoperabilita - sdílené a/nebo referenční slovníky
- Ekosystém - užití standardů RDF, OWL, SPARQL a jejich rozšíření (např. GeoSPARQL)
- dynamický import - URI vždy obsahuje poslední verzi

Dobré praxe - validace schématu SHACL

SHACL

- je graf kompletní?
- obsahuje chyby

Validační schéma tvarů tříd (Shapes) - množina podmínek, které musí třída splňovat

- validace typů
- předpokládané omezení kardinality

validace pro daný cíl

shapes tvoří graf tvarů (jmenný prostor sh http://www.w3.org/ns/shacl#)

Dobré praxe - validace schématu SHACL

Konkrétní postup:

- specifikovat tvary před návrhem ontologie
- navrhnout ontologii a vygenerovat tvary z návrhu
  - potom je ručně zkontrolovat

https://shacl-play.sparna.fr/play/

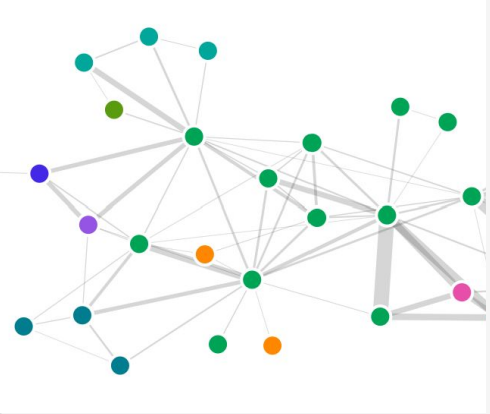
Dobré praxe - reusability?

Ontologie by měly být navrženy tak, aby mohly být použity vícekrát?

Co ale víme o budoucím využití?

Evaluace - jak dobrá je naše ontologie?

jak přesný model světa to je?



https://app.flourish.studio/@flourish/network-graph
https://commons.wikimedia.org/wiki/File:Applications-Earth-Globe.svg

Evaluace - jak dobrá je naše ontologie?

Přesnost (jak model odpovídá expertnímu popisu)

Adaptibilita (můžeme ontologii rozšiřovat bez nutnosti refactoringu?)

Srozumitelnost (je ontologie dokumentovaná? rozumí jí někdo?)

Úplnost (popisuje celou doménu? víme, na které otázky zná ontologie odpovědi?)

Výpočetní složitost (lze použít reasonery? u CDOC-CRM např. ne)

Výstižnost (obsahuje ontologie minimální teorii? obsahuje pouze důležité pojmy?)

Konzistence (obsahuje ontologie kontradikce? odpovídá formální část slovnímu popisu?)

Organizační způsobilost (lze ontologii publikovat? je všechno právně v pořádku?)

Analýza ontologií - metodologie OntoClean

- Identity - Stejnost +I -I (non Identity)
  - kdy jsou nějaké instance tatáž instance?
  - Čím je definován trojúhelník? Délkou tří stran. Tedy všechny takové trojúhelníky jsou identické.
- Unity - Jednota +U -U ~U (anti Unity)
  - vztah mezi instancemi a jejich třídou - jsou instance zároveň částmi, nebo jsou něco jiného?
  - Třídu Hlína lze rozkouskovat na různá množství Hlíny. S třídou Člověk to udělat nelze.
- Rigidity - Neměnnost +R -R ~R (anti Rigidity)
  - příslušnost ke třídě a možnost změny
  - Člověk zůstává týmž člověkem, i když si ostříhá vlasy.
- Dependence - Závislost +D -D
  - požaduje existence jedné třídy existenci jiné třídy?
  - Třída Student nemá smysl bez třídy Učitel.

https://link.springer.com/chapter/10.1007/978-3-540-24750-0_8

Úkoly spojené s ontologickým inženýrstvím

- návrh (Ontology Design)
- kombinování (Ontology Alignment)
- evaluace (Ontology Evaluation)
- automatický návrh ontologie (Ontology Learning)
- (automatické) doplnění instancí (Knowledge Graph Population)

Jak doplnit instance do znalostního grafu

- z textů (zpracování přirozeného jazyka, Natural Language Processing, NLP)
  - rozpoznání entit
  - hyperonyma
  - rozpoznání vztahů
  - synonyma
  - nalezení zobrazení do ontologie
  - hotový software: Open IE https://nlp.stanford.edu/software/openie.html, ChatGPT
- z existujících ontologií
- z databází
  - R2RML https://www.w3.org/TR/r2rml/
- crowdsourcing