# Na obsah zaměřená metoda pro převod tezauru na sémanticky odpovídající ontologie s využitím OWL

*A content-focused method for re-engineering thesauri into semantically adequate ontologies using OWL*

Ludmila Šejvlová

Vytvoření ontologie zachycující velké množství znalostí může být časově velice náročné. Vzhledem k tomu je žádoucí přebírat již existující modely. Z tohoto hlediska jsou zajímavým zdrojem znalostí kontrolované slovníky a to především proto, že se utvářejí desítky let a obsahují několik tisíc nebo dokonce stovky tisíc konceptů a termínů přirozeného jazyka.

Autoři článku se zaměřili na převod slovníků na ontologie popsané pomocí OWL a zároveň odpovídající sémantice deskripční logiky (OWL-DL). Popisovaná metoda se zaměřuje nejenom na vyhovění OWL standardům, ale také na vyvinutí sémanticky odpovídajících ontologií, které:

a. plně využívají sémantické expresivity OWL,

b. jsou konzistentní a poskytují smysluplné výsledky odvozování,

c. mohou být dále integrované s dalšími ontologiemi vytvořenými pomocí stejného postupu.

Autory uváděná metoda se zaměřuje konkrétně na převod tezauru, jakožto speciálního typu slovníku. Tezaurus je typ kontrolovaného a strukturovaného slovníku, jehož strukturní vlastnosti jsou dobře definovány díky mezinárodní standardizační snaze. Jelikož existuje pravděpodobně několik stovek tezaurů, které by mohly být převzaty jako ontologie, tato metoda má potenciál najít široké uplatnění.

Koncepty v tezauru jsou (až na výjimky) entity popsané obecnými termíny, které jsou použitelné u více než jednoho individua. Což znamená, že při převodu tezauru na ontologii skončí většina těchto entit v TBoxu. TBox je komponenta, která popisuje jednotlivé třídy a vazby mezi nimi, zatímco ABox popisuje jednotlivá individua za pomoci definic vytvořených v TBoxu.

OWL sebou přináší některá omezení, často označována jako omezení expresivity, např. OWL kvantifikátory musí být počitatelné. Není možné vyjádřit, že je něčeho hodně, dostatek nebo akorát. Převod tezauru, který tyto možnosti má, do OWL proto není triviální.

Metoda popsaná v článku byla vyvinuta ve dvou fázích, v první vznikl naivní algoritmus pro převod na základě existující literatury a v druhé fázi byl upravován a validován na případových studiích.

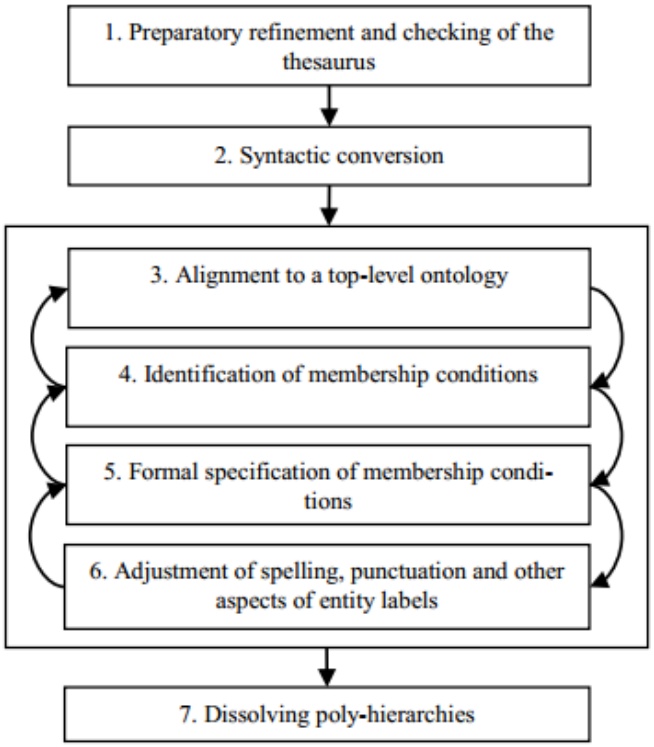
Metoda převodu se skládá ze sedmi kroků, které jsou znázorněny na obrázku 1. Aplikováním níže popsaných kroků se postupně tezaurus transformuje na ontologii.

### Krok 1: Příprava a kontrola tezauru

Ne všechny tezaury dodržují standard, který byl pro ně definovaný. Je potřeba zkontrolovat a případně opravit, pokud tezaurus některé standardy nedodržuje např. nerozlišuje mezi koncepty a termíny.

### Krok 2: Naivní konverze

Krok zaměřený na reprezentaci tezauru ve formálním jazyce OWL-DL, který umožňuje nezaměnitelný výklad vznikající ontologie. Také formální reprezentace v OWL-DL umožňuje automatizovaně ověřovat konzistenci ontologie např. absenci kontradikcí. Jedná se ale stále o naivní konverzi, která tyto chyby obsahovat může, protože v dalších krocích je úkolem tyto chyby odstranit.




Obrázek 1: Kroky metody pro převod tezauru na ontologii

### Krok 3: Zarovnání na top-level ontologii

Hierarchie tezauru může být směs generických a jiných druhů vazeb a proto po druhém kroku, naivní konverzi, mohou vzniknout různé odpojené části tříd organizovaných pomocí „is-a“ vztahů. Pokud bychom to takto ponechali, třídy nejvyšší úrovně budou nejspíše přiřazené k nejobecnější třídě tzv. „Thing“ (jak je nazývána v Protégé). Cílem tohoto kroku je zorganizovat tyto části do souvislé „is-a“ hierarchie a tedy připojit je přímo, nebo nepřímo do společné top-level ontologie.

### Krok 4: Identifikace vzájemných podmínek

Jedinečná výhoda na logice založeném ontologickém jazyku jako je OWL je, že umožňuje specifikovat význam jednotlivých tříd pomocí nezbytných podmínek členství. Cílem tohoto kroku je zajistit co nejkompletnější charakteristiky, které působí jako podmínky členství. Tyto charakteristiky mohou být využity právě v automatické kontrole konzistence. Podmínky členství potom rozhodují o tom, který jedinec patří ke které třídě. V tomto kroku je zapotřebí uskutečnit následující akce:

a. identifikovat význam tříd,

b. vytvořit soubor definic v přirozeném jazyce,

c. rozhodnout, podle jakého principu začlenit třídy a vyřešit dvojznačnosti,

d. vymezit podmínky členství.

### Krok 5: Formální specifikace podmínek členství

Tento krok je zaměřen na vyjádření všech podmínek shromážděných v předchozím kroku do formálního jazyka OWL pomocí společné množiny formálně vymezených vztahů. Vzhledem k tomu že doménové ontologie jsou doplněné do stejné „top-level“ ontologie, vybrané v kroku 3, výsledkem tohoto kroku je stav, kdy vznikající ontologie, top-level ontologie stejně jako další přijaté a nakonec doplněné doménové ontologie jsou integrované.

### Krok 6: Úprava pravopisu, interpunkce a další hlediska popisků entit

Je vybrána jmenná konvence tříd a dalších entit a podle ní jsou všechny popisky upraveny. To má za cíl zlepšit čitelnost jak pro vývojáře, tak pro uživatele. Popisky v ontologii by měly být co nejvíce samovysvětlující a jejich případná změna, nebo úprava nemění sémantiku tříd pro počítače.

### Krok 7: Rozpad poly-hierarchie

Aby byla ontologie snadno udržovatelná měla by se skládat z jediné „is-a“ hierarchie. Tedy poly-hierarchie by měly být rozpadnuty. V kroku 3 se odstraňují sémanticky chybné poly-hierarchie, které vedou k protichůdným členským podmínkám. Tento krok je volitelný, protože sice mění strukturu, ale ne sémantický význam ontologie. Mono-hierarchii je snadnější implementovat a udržovat, ale někdy může být obtížné rozhodnout, který „is-a“ vztah je možné rozpustit.

Metoda popsaná výše v článku se zabývá převodem tezauru na sémanticky odpovídající ontologii. Od jiných existujících metod se liší především tím, že plně využívá OWL pro definování významu jednotlivých konceptů. Hlavní síla metody spočívá v tom, že výsledná ontologie je věrná reprezentace reálných věcí a jednotlivé ontologie se spolu mohou dále kombinovat.

## Zdroje

[1] Daniel KLESS, Ludger JANSEN a Simon MILTON. *A content-focussed method for reengineering thesauri into semantically adequate ontologies* [online]. [cit. 2016-11-26]. Dostupné z: `http://www.semantic-web-journal.net/system/files/swj960.pdf`.