TŘI PŘÍSTUPY K TVORBĚ FORMÁLNÍCH ONTOLOGIÍ

Rostislav Miarka, Alena Lukasová
Katedra informatiky a počítačů, Přírodovědecká fakulta, Ostravská univerzita v Ostravě, 30. dubna 22, 701 03, Ostrava, {rostislav.miarka, alena.lukasova}@osu.cz

ABSTRAKT:
V příspěvku se uvažují tři přístupy k tvorbě formálních ontologií. Prvním je přístup formální konceptové analýzy (Ganter a Wille), založené na matematické formální reprezentaci konceptovými (Galoisovými) svazy. Druhým je konceptově orientovaný datový model (Savinov), který je taktéž reprezentovatelný Galoisovým svazem. Třetím je formální ontologická analýza (Guarino a Welty), která způsob vytváření ontologie přibližuje klasickému pojetí filozofické ontologie. Rozdílnost přístupů je demonstrována na příkladě z literatury.

KLÍČOVÁ SLOVA:
Formální ontologie, formální konceptová analýza, konceprově orientovaný datový model, formání ontologická analýza.

1 Formální ontologie

Pro pojem ontologie existuje více významů. V klasické filozofii se pod pojmem ontologie rozumí teorie bytí. V oblasti umělé inteligence bývá ontologie označována přívlastkem „formální“ a jedná se zde o formalizovaný popis sémantiky objektů určité zájmové domény. Dalším významem pojmu ontologie je, že jako ontologie bývá označován terminologický slovník hierarchicky uspořádaných pojmů z určité problémové domény, který obsahuje i axiomy (pravidla) platné pro tuto doménu. Existuje více přístupů k tvorbě ontologií, zde budou diskutovány tři z nich. Všechny tři uvedené přístupy mají společný rys – konceptově orientované vidění modelované domény. Kromě formálně specifikovaných významů termínů, explicitně vymezující příslušnost ke konceptům, se zpravidla do ontologií zařazují další logické formule, které vyjadřují vztahy mezi koncepty, např. subsumpce, disjunkce konceptů.

2 Formální konceptová analýza

Základy formální konceptové analýzy vytvořili Bernhard Ganter a Rudolf Wille. [1] Formální konceptová analýza je analýzou mnohorozměrných dat, která je vybudovaná na základě Galoisových svazů (označují se jako konceptové svazy). V datové matici jsou v řádcích vektory popisující zkoumané objekty a ve sloupcích vektory atributů pro jednotlivé objekty. Datová matice reprezentuje binární relaci a představuje formální kontext. Formální konceptová analýza má za úkol nalézt ve zkoumaných datech přirozené shluky objektů nebo přirozené shluky atributů. Přirozený shluk objektů je množina všech objektů, které sdílejí nějakou množinu atributů. Přirozený shluk atributů je množina všech atributů, kterými se vyznačuje určitá množina objektů. Množina objektů s jejich atributy, která je zároveň přirozeným shlukem objektů i přirozeným shlukem atributů, se označuje jako koncept.

Příklad 1.
Příklad z literatury podle J. L. Pfalze a Ch. M. Taylora. [9] Je dán kontext R (Tabulka 1), tj. relace mezi množinou atributů A (Tabulka 3) a množinou objektů O (Tabulka 2).

Tabulka 1. Kontext **R**

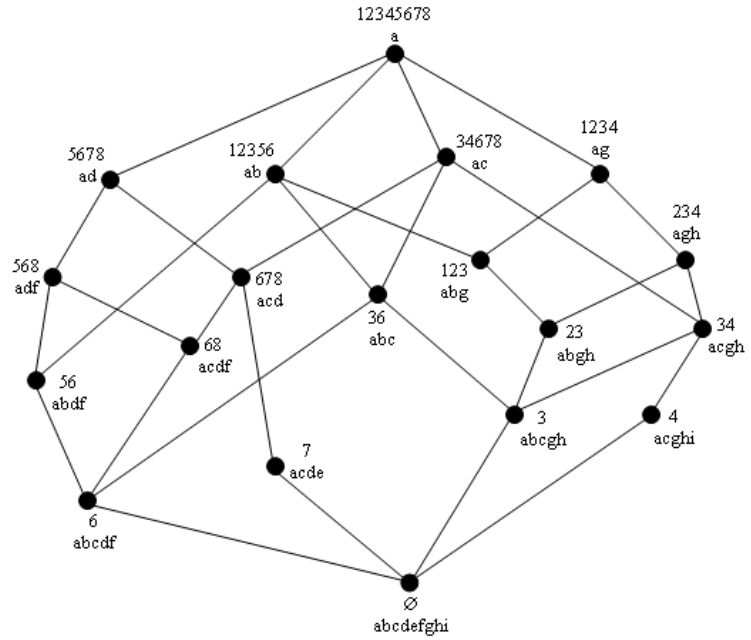
R	a	b	c	d	e	f	g	h	i
1	×	×					×		
2	×	×					×	×	
3	×	×	×				×	×	
4	×		×				×	×	×
5	×	×		×		×			
6	×	×	×	×		×			
7	×		×	×	×				
8	×		×	×		×			

Tabulka 2. Objekty v kontextu **R**

	objekty
1	pijavice
2	cejn
3	žába
4	pes
5	plevel
6	třtina
7	bob
8	kukuřice

Tabulka 3. Atributy v kontextu **R**

	atributy
a	potřebuje k životu vodu
b	žije ve vodě
c	žije na souši
d	potřebuje k tvorbě potravy chlorofyl
e	klíčí dvěma malými lístky
f	klíčí jedním malým lístkem
g	může se pohybovat
h	má páteř
i	kojí potomky



Obr. 1. Svaz kontextu **R** řešený formální konceptovou analýzou

Svaz LR na Obr. 1 je vizuálním modelem obsahu kontextu **R**. Každý koncept začleněný do grafu představuje uzel se dvěma částmi návěští: V horní části návěští je uveden extent konceptu, tj. objekty, které koncept reprezentuje, a v dolní části jeho intent, tj. množina atributů, kterými se koncept vyznačuje. V obrázku je svaz orientovaný vzhledem k O, kde univerzální koncept (top-koncept) T, jehož extent zahrnuje všechny objekty, které zde sdílejí atribut *a*, je supremem svazu, absurdní koncept (bottom-koncept) ⊥ je prázdný koncept vyznačující se všemi atributy *abcdefghi*, je infimem svazu. Konceptu s extentem 123, který se vyznačuje atributy *abg* (potřebující vodu k životu, žijící ve vodě a mající možnost pohybu) určujícími jeho intent lze např. přiřadit jméno „vodní živočich“ jako nový atribut.

Pro orientaci ve výsledném konceptovém svazu orientovaném vzhledem k množině objektů O zde platí pravidlo: Koncept C1 je subkonceptem konceptu C2, právě když existuje cesta grafem směrem dolů od uzlu konceptu C2 k uzlu konceptu C1.

3 Konceptově orientovaný datový model

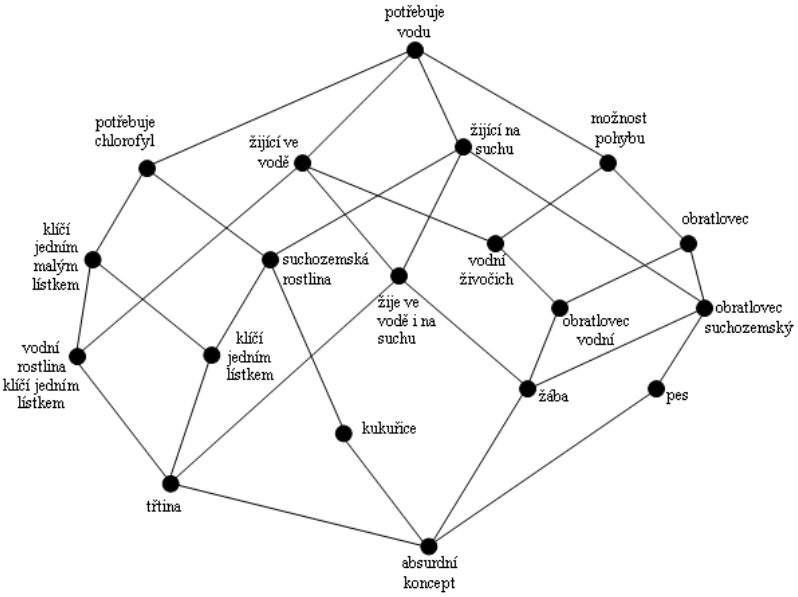
Autorem tohoto přístupu je A. Savinov. [10] Konceptově orientovaný datový model je založený na myšlence, že koncepty, určené prostřednictvím svých instancí (objektů), existují a zároveň tvoří určité prostory, přičemž struktura prostoru (ontologie) popisuje syntax a struktura objektů náležejících konceptu, určená prostřednictvím jejich atributů, reprezentuje sémantiku. Vlastnosti konceptu jsou určeny jeho superkoncepty. Základními prvky prostoru jsou koncepty. Objekty, které jsou instancemi konceptů, jsou reprezentovány datovými elementy. Koncepty jsou analogické relacím, objekty jsou prvky relací v relačním datovém modelu. Zároveň je koncept analogický k třídě v objektově orientovaném modelu. Konceptově orientovaný datový model je multidimenzionální hierarchický model založený na částečně uspořádaných množinách a teorii svazů. Pro matematickou reprezentaci modelu slouží Galoisovy svazy.

Příklad 2.
Stejný příklad, který byl v předchozí kapitole řešen formalní konceptovou analýzou. Do grafu (Obr. 2) byly doplněny atributy určující příslušné koncepty.

Konceptově orientované modelování vychází z předpokladu, že každý uvažovaný atribut je též konceptem, jehož význam je dán jeho extentem. Z hlediska konceptově orientovaného datového modelu jsou tedy všechny atributy tabulky určující kontext **R** pojímány jako koncepty ve vztazích subkoncept - superkoncept. Extent superkonceptu má přitom kardinalitu vždy menší nebo rovnou kardinalitě extentu jeho subkonceptu. Např. extent konceptů žije_na_suchu i žije_ve_vodě je podmnožinou extentu konceptu potřebuje_k_životu_vodu, představující v konceptovém svazu na Obr. 2 jeho top-koncept. Přístup formální konceptové analýzy spočívá v postupných extrakcích výskytů taxonomického vztahu subkoncept ⟶ superkoncept (formálně reprezentovaného vztahem isa) na základě uvažovaných atributů prvků dané domény.
Tyto vzájemné vztahy subkoncept ⟶ superkoncept dávají možnost specifikace relace částečného uspořádání, která je pro modelový konceptový svaz základní.
Přístup opírající se o sémantiku konceptů (jejich extenty) vede k nalezení i takových prvků relace subkoncept ⟶ superkoncept, které formální konceptová analýza neodhalí, neboť nejsou v datech ukryta. Takovým prvkem relace je např. vztah potřebuje_chlorofyl ⟶ žije na suchu.
Další výhodou konceptově orientovaného modelování, opírajícího se o sémantiku konceptů, je možnost pojmenování konceptů. Např. koncept s extenzí 3,6 {žába, třtina} a intenzí a,b,c určenou logickou konjunkcí

potřebuje_k_životu_vodu & žije_ve_vodě & žije_na_suchu

lze pojmenovat jako žije_ve_vodě_i_na_suchu.
Multidimenzionální charakter reprezentujícího konceptového svazu je vidět např. na příkladech vztahů subkoncept ⟶ superkoncept mezi koncepty suchozemská_rostlina ⟶ potřebuje_chlorofyl a suchozemská_rostlina ⟶ žije_na_suchu ⟶ potřebuje_vodu.
Po doplnění atributů určujících příslušné koncepty do grafu z Obr. 1 je zřejmé, že si výstupy obou přístupů v podstatě odpovídají, neboť v následujícím obrázku vytvořeném konceptualizací na základě subsumpce konceptů se zhruba jedná o grafickou reprezentaci konceptově orientovaného datového modelu vytvořeného formální konceptovou analýzou na základě (doplněného) kontextu **R**.



Obr. 2. Grafická reprezentace konceptově orientovaného datového modelu vytvořeného na základě kontextu **R**

4 Formální ontologická analýza

Autoři Guarino a Welty [2] – [8] definují svůj cíl takto: logickými prostředky formalizovat všeobecně sdílené znalosti o daném světě, založené na rigorózní charakterizaci základních ontologických konceptů (kategorií), jako jsou ty, které berou v úvahu prostor, čas a strukturu fyzických objektů modelovaného světa, pomocí nevelkého souboru jejich meta-vlastností. Autoři definují identitu, esenci, jednotu, rigiditu, závislost a další jako meta-vlastnosti sledovaných vlastností, určujících koncepty a role a na jejich základě navrhují metodologii tvorby ontologií. Úkolem formální ontologické analýzy je na základě meta-vlastností určit stanovení významu (pozice), který má uvažovaná vlastnost v rámci určité konceptualizace. Použití meta-vlastností zde nevyžaduje logiku vyšších řádů. Primitivní meta-vlastnosti jsou pojmenovávány jmény začínajícími znaky "+", "–", "¬" a "~". Tyto znaky mají charakter operátorů. ϕ^M znamená, že vlastnost ϕ má meta-vlastnost M. Využívá se též modálních operátorů modální logiky □/◊ (jistě platí/možná platí).

4.1 Definice meta-vlastností a jejich vztahů

U entit modelovaného světa je potřeba sledovat, zda entita zůstane stejná i při změně některých svých vlastností v čase. První meta-vlastností je esencialita. Esenciální vlastnost entity je vlastnost, kterou entita musí mít. Z esenciality vychází pojem rigidity. Všechny rigidní vlastnosti tvoří v ontologii tzv. páteřní taxonomii.

Definice 1. *Rigidní vlastnost +R entity je vlastnost ϕ, která je esenciální pro všechny její instance, tj.*

$$\forall x\,(\phi(x) \rightarrow \Box\phi(x)).$$

Non-rigidní vlastnost –R entity je vlastnost, která není esenciální pro některé instance, tj.

$$\exists x\,(\phi(x)\ \&\ \neg\Box\phi(x)).$$

Anti-rigidní vlastnost ~R entity je vlastnost, která není esenciální pro všechny své instance, tj.

$$\forall x\,(\phi(x) \rightarrow \neg\Box\phi(x)).$$

Částečně-rigidní vlastnost ¬R je vlastnost, která je non-rigidní, ale není anti-rigidní.

V následující definici označuje **E**(*x*,*t*) časově závislý predikát, který znamená že *x* aktuálně existuje v čase *t*. Podmínku identity označujeme IC.

Definice 2. *Podmínka IC je formule Σ, vyjadřující stejnost (nerozlišitelnost), která splňuje (1) nebo (2)*

$$\neg\Box(\mathbf{E}(x,t) \,\&\, \phi(x,t) \,\&\, \mathbf{E}(x,t') \,\&\, \phi(x,t') \,\&\, x = y \rightarrow \Sigma(x,y,t,t')) \qquad \textbf{(1)}$$

$$\neg\Box(\mathbf{E}(x,t) \,\&\, \phi(x,t) \,\&\, \mathbf{E}(x,t') \,\&\, \phi(x,t') \,\&\, \Sigma(x,y,t,t') \rightarrow x = y) \qquad \textbf{(2)}$$

IC je nutná, splňuje-li (1), postačující, splňuje-li (2).

Definice 3. *Vlastnost je nositelkou IC, právě když je pod-vlastností vlastnosti, která poskytuje tuto IC.*

Definice 4. *Vlastnost ϕ poskytuje IC, právě když*

1. *je rigidní,*
2. *existuje pro ni IC,*
3. *tutéž IC nenesou všechny její pod-vlastnosti.*

Definice 5. *Vlastnost, která je nositelkou IC, se nazývá sortal. Vlastnosti nesoucí IC se označují +I, v opačném případě –I. Vlastnost poskytující IC se označují +O, v opačném případě –O.*

Uvedený pojem identity vychází z intuitivní představy, jak agent (člověk, program) rozpoznává individuální entity svého světa. Rozhodování identity závisí na zaměření konceptualizace zájmové domény. Pro ujasnění pojmu identity je důležité si uvědomit rozdíl mezi identitou a jednotou. Pomocí identity rozlišujeme instance dané třídy od jiných instancí této třídy. Jednota se vztahuje k problému rozlišení části instance od zbytku světa pomocí sjednocující relace.

Definice 6. *Objekt **a** je celkem pod ω, právě když ω je relací ekvivalence, která svazuje právě všechny části **a**.*

Definice 7. *Vlastnost ϕ je nositelkou podmínky jednoty (UC), právě když existuje jedinečná relace taková, že každá instance vlastnosti ϕ je nutně celkem pod ω. Vlastnost nesoucí podmínku jednoty je označována +U, v opačném případě –U. Vlastnost ϕ je nositelkou podmínky anti-jednoty, jestliže každá její instance není nutně celkem. Anti-jednotná vlastnost se označuje ~U.*

V následující definici relace **C**(*x*,*y*) označuje součást celku, na rozdíl od **P**(*x*,*y*), označující část celku.

Definice 8. *Vlastnost ϕ je externě závislá na vlastnosti ψ, jestliže pro každou její instanci x musí nutně existovat instance vlastnosti ψ, která není ani část ani součást x:*

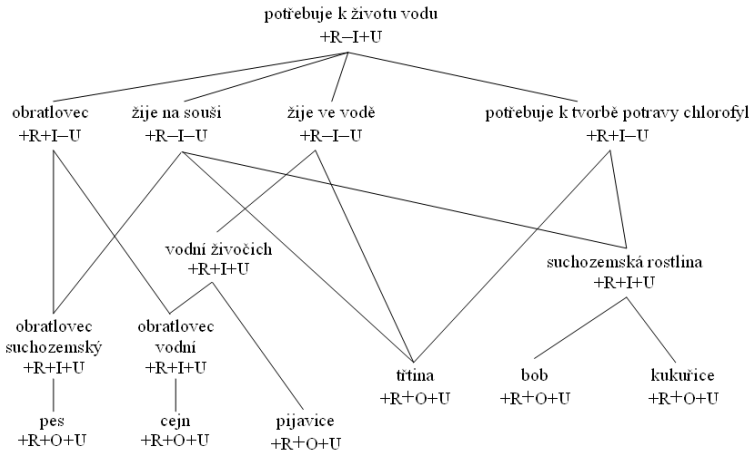
$$\forall x \,\Box(\phi(x) \rightarrow \exists y \,\psi(y) \,\&\, \neg P(y,x) \,\&\, \neg C(y,x)).$$

Externě závislé vlastnosti se označují +D (jinak –D).

Příklad 3.
Stejný příklad z literatury jako v předchozích kapitolách. Nejprve u objektů a atributů z kontextu R určíme meta-vlastnosti. Jak objekty tak atributy zde vystupují jako koncepty – množiny entit, které sdílejí určité vlastnosti. Členění na vlastnosti a objekty je zde pouze kvůli přehlednosti. Mezi vlastnosti jsme přidali i nově pojmenované vlastnosti, které vznikly v konceptově orientovaném datovém modelu.

Vlastnost	rigidita	identita	jednota	Objekt	rigidita	identita	jednota
potřebuje k životu vodu	+R	–I	+U	pijavice	+R	+O	+U
žije ve vodě	+R	–I	–U	cejn	+R	+O	+U
žije na souši	+R	–I	–U	žába	~R	+O	+U
potřebuje k tvorbě potravy chlorofyl	+R	+I	+U	pes	+R	+O	+U
klíčí dvěma malými lístky	~R	+I	+U	plevel	~R	+I	+U
klíčí jedním malým lístkem	~R	+I	+U	třtina	+R	+O	+U
může se pohybovat	~R	+I	+U	bob	+R	+O	+U
obratlovec	+R	+I	+U	kukuřice	+R	+O	+U
kojí potomky	~R	+I	+U				
vodní živočich	+R	+I	+U				
suchozemská rostlina	+R	+I	+U				
obratlovec vodní	+R	+I	+U				
obratlovec suchozemský	+R	+I	+U				

U všech konceptů jsme určili rigiditu, identitu a jednotu. Např. koncept pes je rigidní, protože kdyby určitá entita ztratila vlastnost „být psem“, přestane být tím, čím je. U identity jsme přiřadili +O, protože jednotlivé psy jsme schopni od sebe odlišit a mají svou vlastní identitu. Jednota +U značí, že entity, které jsou psi, tvoří vždy celek. U konceptu žába je identita a jednota stejná. U rigidity je přiřazeno ~R (anti-rigidní). Určitá entita může být žábou, ale stejně tak může být pulcem, i když je to pořád ta samá entita. Takže ne vždy platí, že entita tuto vlastnost musí mít.

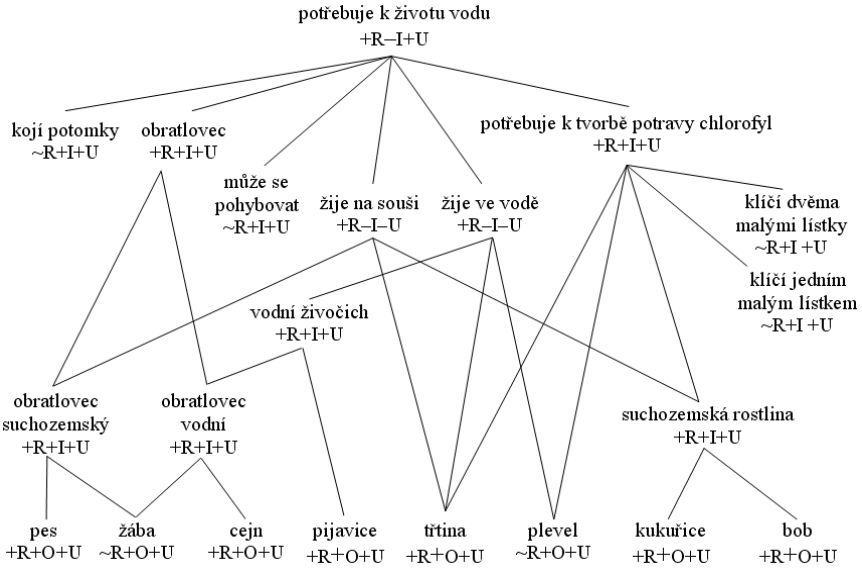


Obr. 3 - Páteřní taxonomie

Páteřní taxonomii tvoří všechny rigidní vlastnosti v ontologii (označené +R). Na základě vztahu subsumpce mezi jednotlivými vlastnostmi a objekty můžeme vytvořit strom reprezentující páteřní taxonomii (Obr. 3). Pokud mezi dvěma koncepty existuje vazba, platí pravidlo, že pro každou instanci podřazeného konceptu platí, že je současně i instancí nadřazeného konceptu. Pro všechny instance konceptu obratlovec vodní platí, že jsou současně instancemi konceptu vodní živočich. Podobně pro všechny instance konceptu kukuřice platí, že jsou instancemi konceptu suchozemská rostlina.

V některých případech mohou v páteřní taxonomii nastat konflikty. Například koncept s meta-vlastností ~U nemůže subsumovat koncept označený +U. V našem případě žádné konflikty nenastaly, páteřní taxonomie je tedy v pořádku. Do taxonomie můžeme zařadit i ostatní (non-rigidní a anti-rigidní) vlastnosti. Ve výsledné taxonomii (Obr. 4) je nutné

odstranit všechny vazby, které by způsobily nějaký konflikt. V našem případě jsou to vazby, kdy ~R subsumuje +R.



Obr. 4 - Výsledná taxonomie

5 Závěr

Cílem tohoto příspěvku bylo seznámit čtenáře s možnostmi využití tří přístupů k tvorbě ontologie dané domény.
Závěrem lze konstatovat:

- Formální konceptová analýza jako analýza dvouhodnotových dat, charakterizujících nějaký kontext v dané doméně, o jejichž významu a struktuře není předem nic známo, je vhodným startovacím krokem k vytvoření ontologie z daného kontextu. Výhodou je zde automatizovatelné provedení analýzy.
- Na základě souboru atributů charakterizujících objekty dané domény a dalších atributů, doplněných po provedení formální konceptové analýzy, lze určit prvky relace superkoncept – subkoncept částečného uspořádání atributů prvků dané domény a uspořádat je do konceptového grafu nastiňujícího výslednou ontologii. K určení vztahů superkoncept – subkoncept je ale potřeba dobrá znalost problémové domény.
- K prověření opodstatnění hierarchického uspořádání konceptů v navrhované ontologii je vhodná metodika Guarina a Weltyho.

Vhledem k tomu, že konceptové svazy vytvořené formální konceptovou analýzou i konceptově orientovaným datovým modelem mají shodný tvar, je možné se mylně domnívat, že jejich výstupem je totéž. Není to ovšem pravda. Formální konceptová analýza vychází z datové matice (formálního kontextu) a na základě dat v ní se pokouší najít skupiny objektů, které sdílejí určitou množinu vlastností. Z těchto skupin je pak vytvořen konceptový svaz. Konceptově orientovaný datový model vychází z konceptů a určení vztahů mezi nimi (subkoncept ⟶ superkoncept). Konceptový svaz je pak vytvořen na základě těchto vztahů. Pro určení vztahů mezi koncepty je potřeba dobře znát problémovou doménu.

Uvedené tři kroky lze považovat za jakousi metodologii tvorby ontologií. Výchozím bodem je formální konceptová analýza, která je pouhou analýzou dat a vůbec nebere v úvahu sémantiku sledovaných objektů a atributů. Přístup konceptově orientovaného datového modelu přiřadí do modelu sémantiku pomocí vztahů subkoncept ⟶ superkoncept. Poslední krok, přístup formální ontologické analýzy (OntoClean), prověří uspořádání konceptů v dané ontologii.

Výstupem tohoto kroku je páteřní taxonomie a celková taxonomie všech konceptů v ontologii. Vazby mezi koncepty jsou zde vytvořeny na základě vztahu subsumpce.

Výhodou uvedené metodologie je, že provedením metody OntoClean se ontologie „vyčistí“ a neobsahuje pak žádné přebytečné vazby. Nevýhodou je, že konceptově orientovaný datový model i formální ontologická analýza vyžadují velmi dobrou znalost problémové domény.

6 Literatura

1. Ganter, B., Wille, R. *Formal Concept Analysis: Mathematical Foundations.* Springer Verlag, Berlin, 2004. ISBN 3-540-62771-5.
2. Guarino, N. (ed.). *Formal Ontology in Information Systems*. IOS-Press, Amsterdam 1998.
3. Guarino, N., Carrara, M., Giaretta, P. An Ontology of Meta-Level Categories. In Doyle, J., Sandewall, E., Torasso, P. (eds.) *Principles of Knowledge representation and Reasoning: proceedings of the fourth International Konference (KR94).* Morgan Kaufman, San Mateo, 1994.
4. Guarino, N., Welty, Ch. A Formal Ontology of Properties. *Proc. of 12th Conf. On Knowledge Engeneering and Knowledge Management.* Lecture Notes on Computer Science, Springer Verlag, 2000.
5. Guarino, N., Welty, Ch. An Overview of OntoClean. In Staab, S., Studer, R. (eds). *Handbook on Ontologies.* Springer, 2004, Springer-Verlag Berlin. ISBN 3-540-40834-7.
6. Guarino, N., Welty, Ch. Conceptual Modeling and Ontological Analysis. *Proc. FOIS 2001 Formal Ontology in Information Systems*, http://www.cs.vassar.edu/faculty/welty/
7. Guarino, N., Welty, Ch. Supporting ontological analysis of taxonomic relationships. In Burkhardt, H., Smith, B. (Eds.). *Handbook of Metaphysics and Ontology.* Philosophia Verlag, 1991, Munich. ISBN 3-884-05080-X
8. Guarino, N. Formal Ontology, Conceptual Analysis and Knowledge representation. In Guarino, N., Poli, R. (eds.): International *Journal on Formal Conceptual Analysis*, 1999.
9. Pfaltz, J.L., Taylor, Ch.M. Scientific Knowledge Discovery through Iterative Transformation of Concept Lattices. *Proc. Workshop of Discrete Math. And Data Mining*, *April 2002*.
10. Savinov, A. Informal Introduction into the Concept-Oriented Data Model. http://conceptoriented.com, 2005.

ABSTRACT:

Three approaches of making formal ontologies

In this paper we discuss three approaches of making formal ontologies. First approach is formal concept analysis (Ganter and Wille) based on mathematic formal representation by concept (Galois's) connections. Second approach is concept-oriented data model (Savinov), which can also be represented by Galois's connection. Third approach is formal ontological analysis (Guarino and Welty), which closes the way of making ontologies to classical approach of philosophical ontology. Difference of these approaches is illustrated on example from literature. Three approaches introduced above can be considered to be a methodology of making formal ontologies. The advantage of this methodology is that after execution of all approaches, the result ontology is clean and it doesn't contain unnecessarily relationships. The disadvantage of this methodology is that it requires a good knowledge in problem domain.