



**MATEMATICKO-FYZIKÁLNÍ  
FAKULTA**  
Univerzita Karlova

## **BAKALÁŘSKÁ PRÁCE**

Dennis Pražák

**Návrh a implementace nástroje na  
vytváření diagramů unifikovaných  
konceptuálních schémat multi-modelových  
a dalších NoSQL databázových systémů  
pomocí prostředků schématických  
kategorií**

Katedra softwarového inženýrství

Vedoucí bakalářské práce: RNDr. Martin Svoboda, Ph.D.

Studijní program: Informatika

Studijní obor: IPP2

Praha 2023

Prohlašuji, že jsem tuto bakalářskou práci vypracoval(a) samostatně a výhradně s použitím citovaných pramenů, literatury a dalších odborných zdrojů. Tato práce nebyla využita k získání jiného nebo stejného titulu.

Beru na vědomí, že se na moji práci vztahují práva a povinnosti vyplývající ze zákona č. 121/2000 Sb., autorského zákona v platném znění, zejména skutečnost, že Univerzita Karlova má právo na uzavření licenční smlouvy o užití této práce jako školního díla podle §60 odst. 1 autorského zákona.

V ..... dne ..... ..

Podpis autora

Děkuji vedoucímu RNDr. Martinu Svobodovi, PhD., za odborné vedení práce a všechny poskytnuté rady a podněty.

Název práce: Návrh a implementace nástroje na vytváření diagramů unifikovaných konceptuálních schémat multi-modelových a dalších NoSQL databázových systémů pomocí prostředků schématických kategorií

Autor: Dennis Pražák

Katedra: Katedra softwarového inženýrství

Vedoucí bakalářské práce: RNDr. Martin Svoboda, Ph.D., Katedra softwarového inženýrství

Abstrakt: Tato práce se zabývá vývojem grafické aplikace pro konceptuální modelování databázových schémat bez předem známého paradigmatu. Dosahuje toho pomocí schématických kategorií, které jsou zobecněním dříve známých modelů jako jsou ER a UML. Tyto modely jsou na schématické kategorie převoditelné a aplikace tuto funkcionalitu umožní. Klasický postup tvorby databázového schématu je nejprve návrh konceptuálního modelu např. ER, který je posléze převeden do logické vrstvy např. na tradiční relační model, který je uplatněn v relačních databázích. V dnešní době je k dispozici mnoho nových modelů a prostředky ER často nestačí k jejich popisu. Výsledná aplikace nabídne uživateli možnost modelovat ve známém schématu ER, automatický převod na schématickou kategorii i možnost modelovat přímo schématickou kategorii.

Klíčová slova: konceptuální modelování, schématická kategorie, databázové systémy

Title: Design and implementation of a tool for diagram creation of unified conceptual schemas of multi-model and other NoSQL database systems using schematic categories

Author: Dennis Pražák

Department: Department of Software Engineering

Supervisor: RNDr. Martin Svoboda, Ph.D., Department of Software Engineering

Abstract: This thesis concerns development of a graphical application for conceptual modeling of database schemas without prior knowledge of the target database paradigm. This is achieved by exploiting schema categories, that are a generalization of previously known models such as ER and UML. These models can be converted into schematic categories and the resulting application will enable this functionality. The traditional procedure for creating a database schema is first the design of a conceptual model using e.g. ER, which is then converted into the logical layer, traditionally the relational model, which is in turn applied in relational databases. Many novel database models are available these days and the resources of ER and UML are often not sufficient to describe them. The resulting application will offer the option to model in the familiar ER schema, automatic conversion to a schematic category, and the option to edit or model the schematic category directly.

Keywords: conceptual modeling, schema category, database systems

# Obsah

<b>1</b>	<b>Teorie</b>	<b>2</b>
1.1	Entity Relationship . . . . .	2
1.2	Schematická kategorie . . . . .	4
	<b>Seznam použité literatury</b>	<b>7</b>
	<b>Seznam obrázků</b>	<b>8</b>
	<b>Seznam tabulek</b>	<b>9</b>

# 1. Teorie

V této kapitole představíme potřebné teoretické koncepty, kterých bude využívat výsledná aplikace. Mezi tyto koncepty patří Entity Relationship (ER), schematická kategorie včetně teorie kategorií a vizuální diagram schematické kategorie.

## 1.1 Entity Relationship

Datový model Entity Relationship (ER) poprvé představil Chen už v roce 1976 [1]. Od té doby se však ER vyvíjel, jak se potřeby datového modelování rozšiřovaly. ER není standardizováno, ale jednu moderní verzi představili Atzeni, Ceri, Paraboschi a Torlone [2, s. 163-179]. Na jejich ER modelu založíme ten náš, který zde popíšeme.

V tabulce 1.1 jsou vyobrazeny jednotlivé konstrukty ER modelu. Zde blíže popíšeme sémantiku jednotlivých konstruktů:

- Entitní typ (Entity Type) reprezentuje entitu. Každá entita má jméno.
- Vztahový typ (Relationship Type) reprezentuje vztah mezi dvěma a více (ne nutně různými) entitami. Každý vztah má jméno.
- Atribut (Attribute) reprezentuje atribut/vlastnost entitních nebo vztahových typů. Každý atribut má jméno.
- Složený atribut (Composite Attribute) je atribut, který má sám atributy. Zakazujeme však další větvení, tedy atributy složeného atributu už samy nemohou být složené. Každý složený atribut má sám jméno, podobně jako jeho vlastní atributy.
- Kardinalita (Cardinality) je dvojice  $(a, b) \in \{0, 1\} \times \{1, *\}$ , kde  $a$  nazýváme minimální kardinalita (spodní hranice) a  $b$  maximální kardinalita (horní hranice). Kardinalitu musí mít všechny atributy a každý účastník vztahu. Výchozí kardinalita je  $(1, 1)$  a ve schématu se většinou neuvádí. Spodní hranice 0 znamená, že účast je volitelná; hranice 1 znamená, že účast je povinná. Horní hranice 1 znamená, že účast je nejvýše jedna; hranice  $*$  znamená, že účastí je libovolný počet.
  - Hranice kardinalit pro jednotlivé účastníky vztahů vyjadřují minimální a resp. maximální počet výskytů jednotlivých instancí účastníků v tomto vztahu.
  - Hranice kardinalit u atributů vyjadřují minimální a resp. maximální počet hodnot atributu, které se vztahují ke každé instanci entity/vztahu.
- Identifikátor (Identifier) umožňuje jednoznačně rozlišit (identifikovat) instance entit. Pro každý entitní typ je povinný alespoň jeden identifikátor, ale může jich být více. Každý identifikátor je tvořen buď
  - jedním nebo více atributy daného entitního typu; takový identifikátor nazýváme interní, nebo
  - jedním, nebo více vztahovými typy, jehož se daná entita účastní, případně kombinací s předchozím; takový identifikátor nazýváme externí.

- Zobecnění (generalization), nebo také ISA hierarchie<sup>1</sup> (ISA hierarchy), vyjadřuje vztah podobný dědičnosti v objektově orientovaném programování. Jde o vztah mezi entitním typem  $E$  zvaným *rodič* a jedním nebo více *děťmi*  $E_1, \dots, E_n$ . Všechny vlastnosti rodiče (atributy, identifikátory, spojené vztahové typy a další ISA hierarchie) jsou i vlastnosti každého z dětí. Každý výskyt dítěte je také výskytem rodiče.

Entitní typy, které nemají ani jeden interní identifikátor (musí mít tedy externí), nazýváme *slabé entitní typy* (weak entity types). Pokud mají interní identifikátor, nazýváme je *silné entitní typy* (strong entity types).

Jednoho vztahu se může účastnit nejvýše jeden slabý entitní typ, protože pro dva takové by byla identifikace nesmyslná. Externí identifikace se však může řetězit, jako na obrázku 1.1. Nesmí ovšem vzniknout cyklus a to ani v kombinaci s ISA hierarchiemi. Formálněji – pokud vytvoříme orientovaný graf  $G = (V, E)$  takový, že

- $V = \text{entitní typy} \cup \text{vztahové typy}$  a
- do hran  $E$  přidáme
  - pro externí identifikátory všechny hrany na orientované cestě od identifikovaného k identifikujícímu entitnímu typu,
  - pro ISA hierarchii pro každý vztah rodič-dítě, kde  $a$  je dítě a  $b$  je rodič, hranu  $(a, b)$ ,

pak graf  $G$  musí být acyklický.

Slabý entitní typ musí být v daném vztahu s kardinalitou  $(1, 1)$ . To proto, že když je instance entity identifikována vztahem, musí tento vztah být jednoznačný – právě jeden.



Obrázek 1.1: Zřetězení externích identifikátorů

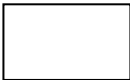
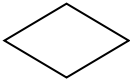

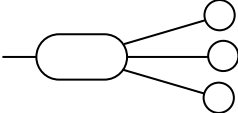
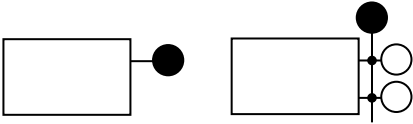
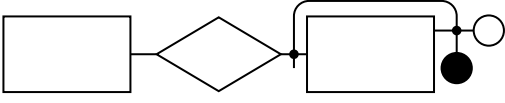
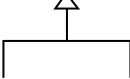
Atributy v ER by měly být pouze elementární vlastnosti. Například pokud by měla mít entita „Zákazník“ atribut „Fyzická adresa“, může být vhodnější fyzickou adresu modelovat jako entitu, neboť má sama atributy jako „Ulice“ a „Město“. Pokud ale víme, že jiná entita nebude v modelu mít fyzickou adresu, můžeme ji případně modelovat jako složený atribut.

U kardinality poznamenáme, že se v ER modelu často dovoluje použít jako hranice libovolná nezáporná celá čísla, tedy  $(a, b) \in \mathbb{N}_0 \times (\mathbb{N}_0 \cup \{*\})$ , tž.  $a \leq b$  (dodfinujeme  $\forall a : a < *$ ). Dají se tak vyjádřit přesnější omezení, např. že jeden uživatel může mít maximálně 5 bankovních účtů. Ve většině případů ale stačí námi definované

<sup>1</sup>ISA z anglického „is a“, analogicky ke vztahu „has a“

hranice kardinality, které vyjadřují volitelnost/povinnost pro spodní hranici a jednočetnost/mnohočetnost pro horní hranici. Toto vymezení nám umožní vyjádřit čtyři nejdůležitější případy, nad kterými se při modelování uvažuje.

Dále upozorníme, že místo  $*$  se v ER modelu může použít symbol  $n$  nebo  $N$  pro vyjádření „libovolného počtu“. Důležitá je ale konzistentnost, aby se v jednom modelu nevyskytovaly dva různé symboly, což by mohlo zmást čtenáře. V této práci budeme používat pouze symbol  $*$ .

Konstrukt	Grafická reprezentace
Entitní typ	
Vztahový typ	
Atribut	
Složený atribut	
Interní identifikátor	
Externí identifikátor	
Zobecnění	

Tabulka 1.1: Grafická reprezentace konstruktů ER modelu, upraveno a přeloženo [2, s. 164]

## 1.2 Schematická kategorie

Nejdříve popíšeme kategorii z teorie kategorií, na níž je schematická kategorie založena.

Kategorie je matematická struktura, která zobecňuje ostatní struktury. Umožňuje mimo jiné studovat vztahy mezi nimi. Poprvé byla představena Eilenbergem a MacLanem v roce 1945 [3].

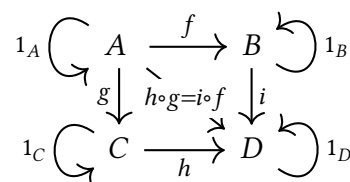
Kategorie  $C = (\mathcal{O}, \mathcal{M}, \circ)$  se skládá z

- třídy objektů  $\mathcal{O}$ ,
- třídy morfismů  $\mathcal{M}$ ; každý morfismus  $f \in \mathcal{M}$  má zdrojový objekt  $A \in \mathcal{O}$ , cílový objekt  $B \in \mathcal{O}$  a říkáme  $f : A \rightarrow B$  ( $f$  je morfismus z  $A$  do  $B$ ),
- operace skládání  $\circ : \mathcal{M} \times \mathcal{M} \rightarrow \mathcal{M}$ ; pro každé dva morfismy  $f : A \rightarrow B, g : B \rightarrow C$  musí  $g \circ f \in \mathcal{M}$  (tranzitivita); pro tuto operaci navíc platí axiomy:

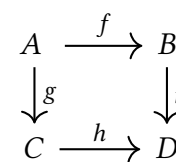


- asociativita – pro morfismy  $f : A \rightarrow B, g : B \rightarrow C, h : C \rightarrow D$  platí  $h \circ (g \circ f) = (h \circ g) \circ f$ ,
- identita – pro každý objekt  $A$  existuje identita  $1_A$ , tž.  $f \circ 1_A = f = 1_B \circ f$  pro každý morfismus  $f : A \rightarrow B$ .

Kategorie je často vizuálně reprezentována schématem připomínajícím orientovaný graf, kde vrcholy jsou objekty a hrany morfismy. Příklad této vizualizace je na obrázku 1.2a. Jedná se o kategorii se čtyřmi objekty  $A, B, C, D$ . V této práci vizualizaci kvůli schematické kategorii budeme zjednodušovat tak, že v ní vynecháme určité morfismy, které budeme považovat za implicitní. Těmito morfismy jsou takové, které jsou v kategorii jen kvůli tranzitivitě, a takové, které jsou v kategorii jen kvůli splnění axiomu identity. Někdy budou tyto morfismy důležité, v takovém případě je v obrázku explicitně uvedeme. Zjednodušená vizualizace je na obrázku 1.2b. Všimněme si, že identity (např.  $1_A$ ) a složené morfismy (např.  $i \circ f$ ) ve schématu nejsou, tedy jsou implicitní.



(a) Se všemi morfismy



(b) Bez tranzitivních a identitních morfismů

Obrázek 1.2: Příklad kategorie

Schematická kategorie je zobecnění databázových schémat založená na teorii kategorií. Jedná se o kategorii, jejíž objekty odpovídají jednotlivým entitním typům, atributům a vztahovým typům ER schématu. Morfismy odpovídají vztahům mezi těmito objekty, přičemž aby byly splněny axiomy kategorie, musí se přidat navíc identity a tranzitivní uzávěr těchto morfismů. Tento koncept společně s algoritmem převodu z ER schématu do schematické kategorie uvádí Martin Svoboda, Pavel Čontoš a Irena Holubová [4]. Pro naše účely se v této práci nebudeme řídit přesně podle této originální publikace, nýbrž schematickou kategorii lehce upravíme.

Schematická kategorie se tedy skládá z třídy objektů a třídy morfismů.

Objekt je čtveřice (identita, název, data  $D$ , množina identifikátorů  $I$ ). Identita je libovolný symbol (např. z  $\mathbb{N}$ ), který rozlišuje a unikátně identifikuje objekty, které mají ostatní složky totožné. Název popisuje textovým řetězcem o jaký objekt se jedná a je zde pro uživatele (čtenáře schematické kategorie). Data je množina tzv. *properties* a platí  $I \subseteq \mathcal{P}(D)$ , tedy je to nadmnožina jednotlivých properties zmíněných v identifikátorech. Každý identifikátor sestává z properties a množina všech identifikátorů je pak množina sestávající z identifikátorů daného objektu. Každý objekt musí mít nejméně jeden identifikátor. Musí platit  $D \supseteq \bigcup I$ . Pro ilustraci – při převodu z ER bude pro původní entitní typy a atributy platit  $D = \bigcup I$ , ale u vztahových typů může být v  $D$  něco navíc.

Morfismus je osmice (signatura, doména, kodoména, směr, název, kardinalita, duplicita, uspořádání). Signatura vyjadřuje „cestu“. Jedná se o řetězec složený ze signatur jednotlivých morfismů, z kterých se morfismus skládá. Pokud se z dalších morfismů

neskládá, je signatura unikátní identitou morfismu (symbol). Pro ilustraci viz obrázek 1.3.

Doména a kodoména jsou identity příslušných objektů, odpovídají tak zdrojovému a cílovému objektu kategorie.

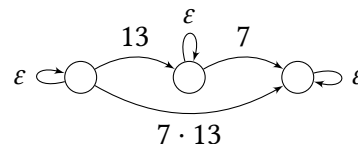
Směr udává směr morfismu (který nemusí nutně odpovídat dvojici doména/kodoména). To proto, že pro každý morfismus bude existovat jeho protějšek. Směr má dvě možné hodnoty a budeme je značit 1 (tam), 0 (zpět).

Kardinalita odpovídá tomu, jak byla popsána pro ER v sekci 1.1 na straně 2. Při skládání se kardinality transformují následovně: pro kardinality  $(a, b)$  a  $(c, d)$  je jejich složením  $(\min(a, c), \max(b, d))$ . Uspořádání nad symboly kardinalit je intuitivně  $0 < 1 < *$ .

Duplicity a uspořádání jsou booleovské hodnoty (`true/false`). Mají význam pouze pokud je maximální kardinalita  $*$ . Říkají, zda-li jsou v instanci modelu u objektů spojených daným morfismem povoleny duplicitní objekty, respektive jestli u nich záleží na pořadí.

Morfismy ve schematické kategorii rozdělíme na několik druhů, které jsou všechny vidět na obrázku 1.3:

- *bázové* (base) morfismy jsou ty, které odpovídají jednotlivým spojením a ISA hierarchiím mezi objekty; signatura je identita,
- *identitní* (identity) morfismy jsou ty, které vznikly kvůli axiomu identity; signatura je  $\varepsilon$ ,
- *odvozené* (composite) morfismy jsou ty, které vznikly kvůli tranzitivitě; signatura je opravdová cesta (zřetěžené signatury); zřetězování zapisujeme ve stejném pořadí jako skládání morfismů, aby bylo analogické.



Obrázek 1.3: Signatury morfismů,  $\cdot$  je operace konkatenace (zřetězování) symbolů

# Seznam použité literatury

- [1] Peter Pin-Shan Chen. The Entity-Relationship Model—Toward a Unified View of Data. *ACM Transactions on Database Systems*, 1(1):9–36, březen 1976.  
doi : 10 . 1145 / 320434 . 320440 .
- [2] Paolo Atzeni, Stefano Ceri, Stefano Paraboschi, and Riccardo Torlone. *Database systems: concepts, languages & architectures*. McGraw-Hill, New York, 1999.
- [3] Samuel Eilenberg and Saunders MacLane. General theory of natural equivalences. *Transactions of the American Mathematical Society*, 58(0):231–294, 1945. URL: <https://www.ams.org/tran/1945-058-00/S0002-9947-1945-0013131-6/>,  
doi : 10 . 1090 / S0002 - 9947 - 1945 - 0013131 - 6 .
- [4] Martin Svoboda, Pavel Čontoš, and Irena Holubová. Categorical Modeling of Multi-model Data: One Model to Rule Them All. In Christian Attiogbé and Sadok Ben Yahia, editors, *Model and Data Engineering*, Lecture Notes in Computer Science, pages 190–198, Cham, 2021. Springer International Publishing. doi : 10 . 1007 / 978 - 3 - 030 - 78428 - 7 \_ 15 .

# Seznam obrázků

1.1	Zřetězení externích identifikátorů . . . . .	3
1.2	Příklad kategorie . . . . .	5
1.3	Signatury morfismů, $\cdot$ je operace konkaténace (zřetězování) symbolů	6

# Seznam tabulek

1.1	Grafická reprezentace konstruktů ER modelu . . . . .	4
-----	--	---