Univerzita Karlova
Filozofická fakulta
Ústav informačních studií a knihovnictví

TEZE DIZERTAČNÍ PRÁCE

PhDr. Ivan Bartoš
ibartos@seznam.cz

Název dizertační práce:

Metodologie a problémy při transformaci dat a určení jejich významu v rámci integrace heterogenních informačních zdrojů

Studijní program:	Informační studia a knihovnictví
Studijní obor:	Informační věda
Forma studia:	kombinované
Vedoucí dizertační práce:	PhDr. Richard Papík, Ph.D.
Předpokládané podání:	únor 2012

OBSAH

Předmluva

Není potřeba zdůrazňovat roli a důležitost informace ve společnosti. Pokud se oprostíme od pojetí informace jako základní lidské potřeby, která následně umožňuje jedinci svobodně se rozhodovat, faktem zůstává, opomineme-li touhu jedince po vědění, která však může být často ve svém důsledku opět kvantifikována, že je dnes informace chápána převážně jako obchodovatelná komodita. Ať se jedná o průmyslové informace, které ovlivňují samotné výrobní procesy, informace o trzích, které v rámci competitive inteligence umožňují konkurenční boj mezi firmami, anebo takové informace, které jsou po důkladné analýze a vložení do obchodního modelu používány marketingovými týmy při rozhodování o další strategii firmy (analýzy odbytu, chování zákazníka/uživatele, atp.) (PAPÍK, 2001).

Každý subjekt shromažďující a skladující informace používá určité, často vlastní, metodologie zpracování znalostí, způsob jejich uložení v relačních nebo jiných strukturách a způsob interpretace uložených informací pro své potřeby nebo pro potřeby jejich obchodování. Tyto metodologie, pokud v daném případě existují, mohou být více či méně korektní, konzistentní a zdokumentované. Míra kvality dat se může negativně odrážet na funkci samotného „mateřského“ subjektu a tím na ziscích, které je schopen generovat (Aberdeen Group, 2007). Výpovědní hodnota informace, její kvalita a kvalita dat, ze kterých vychází, se v různých systémech liší. Toto se děje nejen z důvodu odlišné typologie určitého zdroje informací, ale často i díky samotném způsobu chápání či zachycení informace o popisované entitě skutečného světa. Podobné systémy, v dizertační práci konkrétně databázové systémy, mohou bezchybně fungovat jako samostatné celky (systém je nějak nastaven a případné chyby, pokud je jejich výskyt zdokumentován, se stávají standardem systému). Veliký problém nastává až v momentě potřeby integrace dvou takových heterogenních systémů a následné migraci informací mezi nimi.

Je třeba si uvědomit, že se nejedná o nějaké výjimečné případy. Hýbe-li světem fenomén nazývaný globalizace, pak je integrace informačních systémů ve světě informačních technologií tím samým fenoménem. Oba tyto fenomény jsou na sobě přímo

závislé, navzájem se evokují a naopak, každý z nich řeší důsledky projevů toho druhého. Proto je popisovaná problematika nejen mimořádně aktuální, ale lze přepokládat, že poptávka po kvalitním zvládnutí metodologie integrace a transformace a jejím praktickém využití bude v budoucnu dále narůstat. Na trhu IT služeb existují již delší dobu subjekty, které se zabývají téměř výhradně teoretickou i technologickou realizací datové integrace (např. Wipro, Informatica, a další.)

Dizertační práce, jejíž teze jsou zde předloženy, bude popisovat právě problematiku integrace systému, jehož navržení, správa, údržba a dokumentace zaostává, jehož kvalita dat a možnost následné interpretace informací je problematická, a jehož tvůrci či administrátoři jsou vzhledem k častému fenoménu nucené integrace (vycházející z akvizic firem) značně laxní k jakékoliv spolupráci. Podobná situace, jež nemusí být samozřejmě situace nejčastější, je ideální výzvou a zároveň příkladem, na kterém lze ilustrovat jednotlivé kroky, které jsou potřebné k dosažení maximální integrace zdrojového systému do cílového systému, a to takovým způsobem, aby cílová homogenita a kvalita dat zůstala zachována. Dizertační práce si bude klást za cíl nabídnout v praxi využitelné konkrétní metody integrace různých zdrojů dat a nastínit možná úskalí, kterých by se měl případný tým zodpovědný za realizaci takovéto integrace vyvarovat. Kromě analýzy systému, jeho struktur a dat v něm uložených, se zaměřím také na metody modelování, tvorby ontologií a mapování vzniklých modelů za účelem automatické integrace systémů a transformace dat v nich obsažených.

I přesto, že se v dizertační práci hodlám zabývat příklady z komerčního prostředí, budou popisované metodologie univerzální i pro aplikace v akademickém či státním sektoru.

I. Analýza a klasifikace kvality dat – Data profilig

<u>Teze 1.</u> Řada společností si teprve nyní uvědomuje, jak málo pozornosti věnovala kvalitě dat v průběhu mnohaletého vývoje svých systémů. Díky přístupu řídícího se principem "Time to market[1]" vzniká řada dočasných často nekonzistentních řešení, která mají za následek nekvalitní data ve zdrojových systémech a s tím problematickou interpretaci informací v nich uložených. Znalost kvality a typologie zdrojových dat je pak z pohledu jejich transformace do jiného systému klíčová.

V práci vydefinuji nejčastější chyby v kvalitě dat. Jsou to zejména tyto:

- nekorektní data
- nepřesná data
- data, která nesplňují byznys pravidla
- nekonzistentní data
- nekompletní data
- neintegrovaná data

Zaměřím se na jejich typologii a důvody jejich vzniku, viz následující například:

Data, která nesplňují byznys pravidla – Jedním z problémů kvality dat je jejich nekonzistentnost vzhledem k obchodním pravidlům, nebo, chceme-li, s jejich logickou podstatou. Jednoduchým příkladem takové chyby může být záznam o uživateli, který má datum vytvoření objednávky dřívější než datum své první registrace do systému. Tato chyba vzniká nejčastěji nesprávnou synchronizací aplikací či modulů, které zajišťují dva různé byznys procesy a) registrace uživatele v nějakém **CRM**[2] systému b) vytvoření objednávky v elektronickém obchodě. Nový uživatel vstupuje do systému (registruje se) za účelem vytvoření své první objednávky. Tyto dva kroky nemají větší časovou dilataci.

[1] **Time to market - (TTM)** je čas, který uplyne od doby, kdy byl produkt navržen do momentu jeho komerčního spuštění či uvolnění do prodeje. Jedná se o dobu od nápadu a rozhodnutí o jeho realizaci do momentu úplné realizace.

[2] **CRM** - Customer relationship management (řízení vztahů se zákazníky) je databázovou technologií podporovaný proces shromažďování, zpracování a využití informací o zákaznících firmy.

Při integraci výstupů těchto dvou procesů musí být brána v potaz jejich logická časová souslednost.

Teze 2. **Pokud chceme pracovat s důvěryhodnými informacemi, je nezbytné dosáhnout nejvyšší možné úrovně datové kvality. Tuto úroveň lze určit pouze standardizovanou datovou analýzou. Analytické techniky aplikované na data vypovídají o správnosti jejich struktury, obsahu a kvalitě. Jejich aplikací, lze zjistit chyby, které systém obsahuje a navrhnout kroky k jejich nápravě.**

Popíši standardní teoretické kroky datové analýzy a její aplikace tj.:

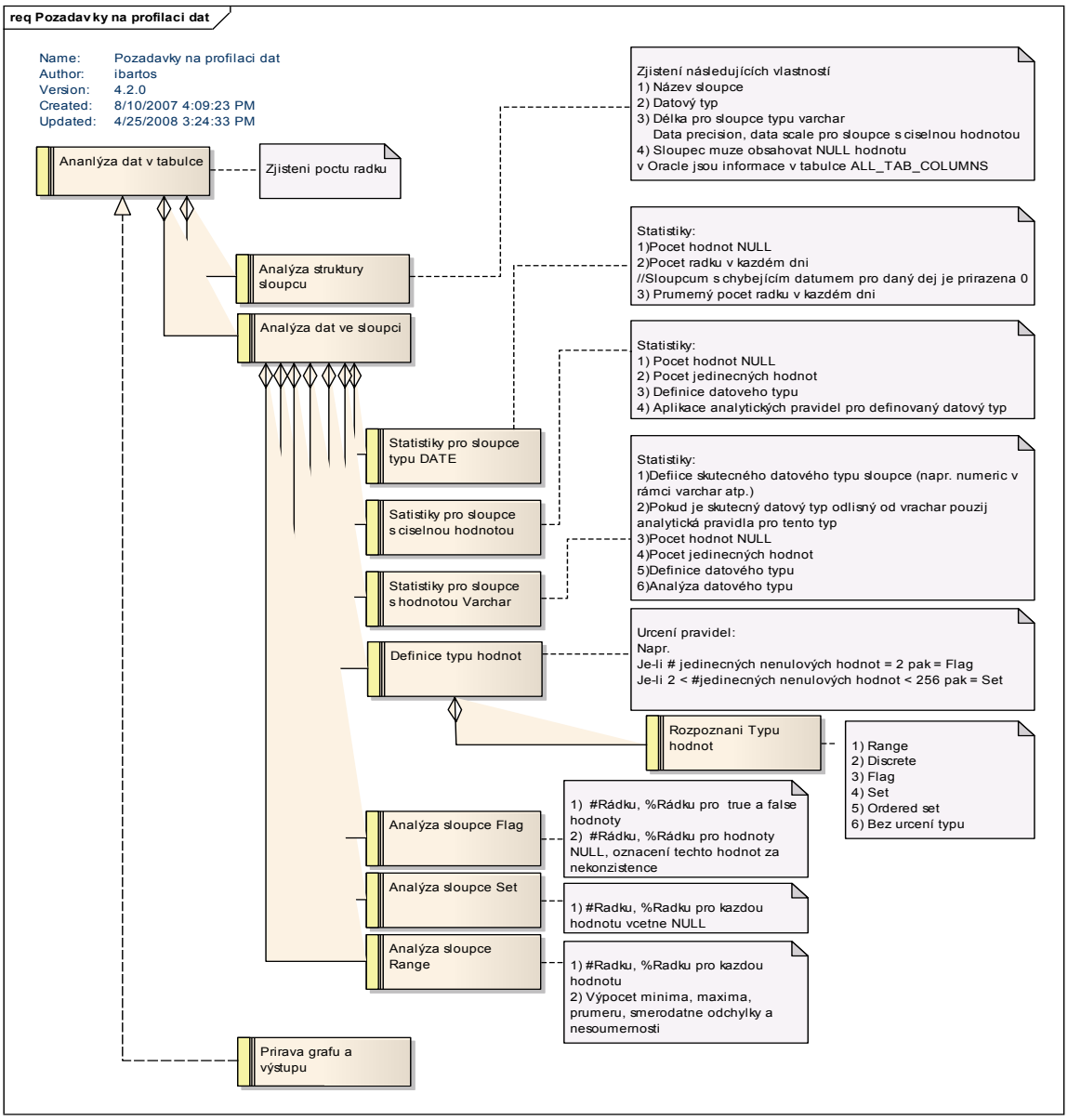
- generování (objevování) pravidel
  (předpokládá existenci mechanismu, který dokáže odhadnout pravidlo, aniž by bylo předem definováno)
- validace pravidel
  (porovnání objevených vlastností s metadaty a případné spolupráci s byznys analytikem)
- validace samotných dat, tedy testování dat oproti objeveným pravidlům
  (výstupem je dokument obsahující informace o potenciálně nekorektních datech)

Teze 3. **Analýza zdrojových dat může probíhat na více úrovních. Samotný systém uchovávající data (DBMS[3]) si určitá metadata schraňuje nativně a lze s nimi proto v analýze pracovat. Tato jsou však pro kvalitní analýzu nepostačující a je nutné je obohatit o sémantickou a syntaktickou rovinu. Takovouto analýzu je vhodné automatizovat z důvodu možnosti její opakované aplikace.**

V rámci dizertace předvedu jednoduchý způsob excerpce metadat ze systémových tabulek či pohledů.

[3] **DBMS -** Database management system – komplexní množina softwarových programů, která řídí organizaci, uložení, správu a získávání dat z databáze. DBMS obsahuje modelovací jazyk pro schéma databáze, datové struktury (pole, záznamy, soubory, objekty), dotazovací jazyk a transakční mechanismy zajišťující integritu dat.

V druhém kroku pak navrhnu jednoduchý použitelný způsob profilace dat nad rámec systémem uchovávaných metadat. Zaměřím se nejdříve na definici požadavků, které budou na výslednou aplikaci (pravděpodobně uživatelský skript), kladeny (viz. následující ilustrace).



Výsledkem takového rozboru požadavků a následně pravidel specifických pro data uložená v databázi bude, po převedení do kódu, aplikace resp. programová jednotka, která dokáže, pokud možno automaticky, provést profilaci dat na požadované úrovni, tj. vyprodukovat potřebná metadata.

Konkrétní programový balíček (package) navrhnu v jazyce PL/SQL pro prostředí Oracle včetně patřičných komentářů v kódu.

Teze 4. **Výsledky analýzy a validace dat jsou často interpretovány na programátorské úrovni. Jako takové jsou však nekomunikovatelné netechnickým osobám, které mají následně zhodnotit, zda je datová kvalita postačující, nebo určovat potřebné kroky k jejímu zlepšení.**

V práci bych se chtěl zaměřit na způsoby uložení a výstupy analýzy dat. Výstupy automatické analýzy jsou vhodné pro další strojové zpracování. Pokud jsou navíc rozšířeny o dimenzi času, kdy jsou jednotlivé statistiky verzovány a trvale uchovávány, stávají se vhodným nástrojem pro monitorování kvality dat v systému v dlouhodobém časovém horizontu (RUSSOM, 2007).

Jako výstup, který je možný předložit jinému oddělení (byznys, oddělení **quality assurance**[4] atp.), jsou však zcela nevhodné a nelze očekávat, že by další strana souhlasila s jejich užíváním. Proto je na místě vybudovat nad těmito zdroji metadat další, pro člověka srozumitelnější rozhraní (BOHUSLAV, 2006).

Navrhnu možné řešení a standardizovaný výstup, který odpovídá požadavkům netechnicky orientovaných oddělení, která se výsledky analýzy dále zabývají.

Teze 5. **Z hlediska analýzy dat nebo statistické analýzy se může jevit objevování pravidel v původním systému jako postačující. V rámci validace dat lze ovšem aplikovat i jiné přístupy. Jedním z možných řešení může být použití již definovaných pravidel (tuto možnost ovšem z důvodu zaměření dizertační práce vypustím), a/nebo jiných vnějších validačních pravidel.**

[4] **Quality assurance -** (zkráceně QA) se týká obecně všeho od návrhu, vývoje, nasazení, údržby až po dokumentaci produktu. Cílem této aktivity je dohlédnout, že výstupy z jednotlivých částí budou mít odpovídající kvalitu, která byla určena.

Jako příklad v práci uvedu aplikování validace a aplikace pravidel za využití regulárních výrazů[5]. Toto předvedu na příkladu emailové adresy dle standardů vycházejících z definice RFC822. Opět zde navrhnu funkční validační balíček a předvedu možné postupy, kterými se lze ubírat. Výsledek pak bude odpovídat navrženému standardu (viz. Teze 4.).

[5] **Regulární výrazy -** Regular Expressions (zkráceně REGEX) - speciální řetězce znaků, které představují určitý vzor (masku) pro textové řetězce. Užívané v C#, Java, Visual Basic .NET, Perl, PHP, Javascript, Unix/Linux, SQL atp.

Transformace informací ze zdrojového do cílového systému

<u>Teze 6.</u> Na základě profilace dat v kombinaci se sémantickou analýzou názvu sloupců a potvrzení významu atributů je možné určit, co která data v původním systému reprezentují. Data jsou klasifikována a mění se v informaci. Na takto klasifikovaná data lze následně aplikovat další metodologie resp. techniky, které je "přetransformují" do informačního konceptu a datové struktury cílového systému.

V práci popíšu metody, které lze se v celém procesu používat. Jednotlivé metody mohou být v rámci různých pojetí charakterizovány a děleny i jiným způsobem. Mnou popisované však tvoří kompletní sadu, která je pro úspěšnou transformaci žádoucí.

Jednotlivé metody budou:

- Extrakce
- Čištění a její algoritmy:
    - Rozpoznání
    - Standardizace
    - Obohacení
    - Unifikace
    - De-duplikace
    - Identifikace (KYJONKA, 2006).
- Oprava
- Konverze
- Transformace

<u>Teze 7.</u> Metodologii ETL[6] procesu lze při přenosu dat aplikovat na libovolný datový prvek, tj. atribut či sadu atributů libovolné entity.

Tyto budu demonstrovat na konkrétním případě aplikace včetně identifikace rizik s nimi spojených.

[6] **ETL -** Extract, transform a load (extrakce, transformace a nahrání) je označení procesu pro přesun dat mezi databázemi. Nejčastěji je používán v prostředí datových skladů.

II. Model entit a vznik datového slovníku

<u>Teze 8.</u> V rámci popisu existujícího systému a struktury dat v něm uložených není vhodné pracovat na úrovní konkrétní implementace (databáze, tabulky, atributy s určitými vlastnostmi). Pro popis je proto vhodné zvolit nějakou vyšší úroveň abstrakce, která definuje systém např. ve smyslu entit, jejich vzájemných vazeb a jejich atributů. Vyšší úrovně abstrakce umožňují intuitivní práci při zkoumání systému i jeho případné integrace. Navržení entit i celého konceptuálního modelu by mělo být prerekvizitou samotného vzniku informačního systému.

V práci vysvětlím, jakým způsobem lze modelovat jednotlivé entity. Vysvětlím architekturu P3A a související modely:

- konceptuální datové modely
- modely logické
- modely fyzické

Zaměřím se převážně na její nejvyšší úroveň tj. konceptuální model a principy jeho budování.

Budu zvažovat volbu mezi matematicko-logickým přístupem a ontologickým přístupem.

Matematicko-logický přístup vychází z teorie množin, relací a matematické logiky,
Ontologický přístup - novější přístup vycházející z paradigmatu ontologií (MOHANEC, 2004)

<u>Teze 9.</u> Ontologie je inženýrská disciplína vhodná jako východisko pro formální teorii konceptuálního modelování.

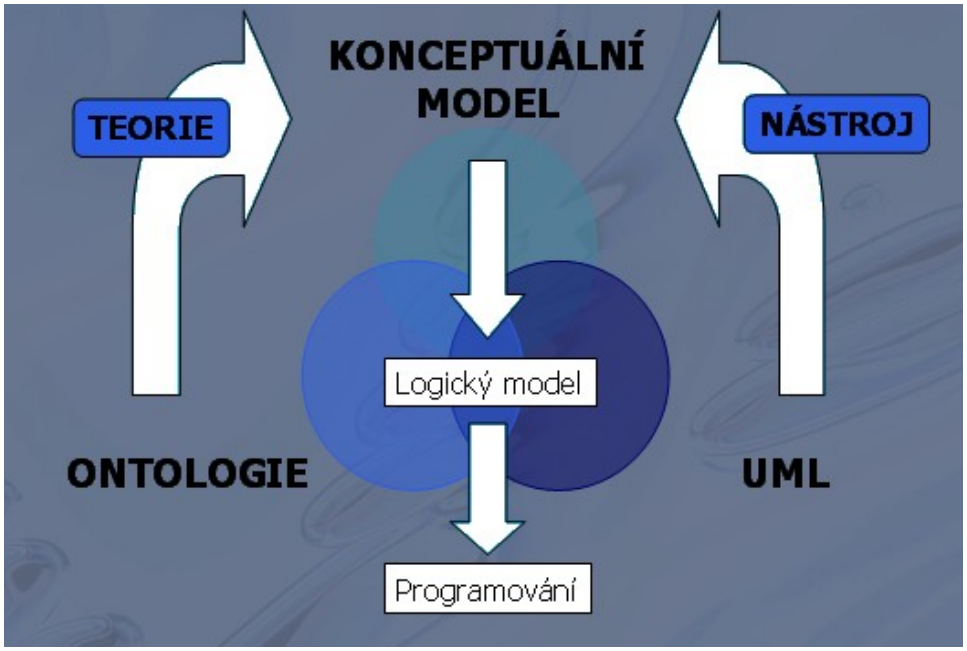
Upozorním na následující skutečnosti:

- Ontologie je univerzální soustava znalostí popisující objekty, jevy a zákonitosti světa „tak jak je“, tj. maximálně nezávisle na lidském usuzování o něm (SVÁTEK, 2002).

- Ontologie je poměrně populární pojem, zvláště v poslední době, kdy vzniká celá řada prací na téma využití ontologií v prostředí Internetu. Hodlám se však zabývat ontologií v jejím původnějším významu, tj. jako základu pro vybudování konceptuálního modelu, který je odrazem reálného světa. Ontologický přístup zkoumá význam jednotlivých pojmů, co je to objekt, co je vlastnost, co je to vztah. Na rozdíl od matematicko-logických přístupů je pro ontologický přístup prvotní význam sám o sobě, a teprve na druhém místě matematicko-logický formalizmus.

V práci rozeberu základní specifikace ontologie, které nesmíme v případě konceptuálního modelování opomenout.

Užití ontologie pro tvorbu konceptuálního modelu definuji takto:



- Ontologie dává přesný význam konceptuálnímu modelu
  - Definuje přesně jeho jednotlivé pojmy
- Konceptuální model je možné vyjádřit v **UML**[7]
  - Používá UML přesně definovaným způsobem
  - Upřesňuje jeho sémantiku s ohledem na použitou ontologii

Následně definuji význam ontologie pro konceptuální modelování:

- Ontologie poskytuje vědecký (filosofický) základ pro výklad konstruktů konceptuálního modelu.

- Společně s logikou poskytuje základ pro formální popis konceptuálního modelu.

Na zvoleném příkladu pak demonstruji využití jazyka OWL (Web Ontology Language), kterým je ontologii možné zapsat např. v nástroji open source Protege – OWL (http://protege.stanford.edu/) a vytvořím grafickou interpretaci ontologie v diagramu.

[7] **UML** - Unified Modeling Language – standardizovaný jazyk vizuální specifikace určený pro modelování objektů. Tento obecný jazyk umožňuje graficky znázornit abstraktní model systému tj. tzv. UML model.

<u>Teze 10.</u> Konceptuální model umožňuje formální reprezentaci dat, která figurují v procesu reálného světa, a na kterých je možné postavit potřebnou byznys aktivitu. Konceptuální model často obsahuje objekty, které ve skutečnosti neexistují, nebo zatím nejsou implementovány např. ve fyzické databázi. Výhoda tohoto modelu je jeho implementační (platformová) nezávislost a s ním související flexibilita v reakcích na změnu.

Zdůrazním, že:

Na konceptuální úrovni se při tvorbě modelu zaměřujeme pouze na identifikaci nejdůležitějších vztahů mezi různými entitami. Takový model může být obecný až do té úrovně, že jeho implementace nemusí záviset na moderních informačních technologiích. Jeho implementaci si lze představit například i jako papírovou kartotéku.

Na konkrétním případě demonstruji kombinaci konceptuálního modelu a logického modelu. Vztahy mezi jednotlivými entitami lze znázorňovat na základě několika možných principů. Na této úrovni abstrakce se jeví nejvhodnější použití rozpoznávání typů vztahů: dědičnost a celek-část, podobně jako je to obvyklé v objektově orientovaném modelování (OOM) (MOHANEC, 2004).

<u>Teze 11.</u> Na základě konceptuálního modelu vzniká obecný datový slovník, který bývá označován jako Enterprise Data Dictionary (EDD), z pohledu dat pak jako meta-metadatový slovník. V něm je zastoupen každý datový prvek obsažený v systému. Existence EDD je pro funkční a konzistentní informační systém nezbytností.

Zaměřím se na vlastnosti datového slovníku:

Datový slovník by měl být základem pro tvorbu jakéhokoliv informačního systému. Datový slovník poskytuje přizpůsobitelnou "ukládací" oblast, nezávislou na aplikaci, ke které se pojí, kde můžeme vytvářet množiny rozšířených atributů položek, které popisují

obsah a vzhled dat, jejichž referencí je pak zajištěna jejich konzistence v samotné implementaci.

Východiskem pro návrh struktury obecného datového slovníků je norma ISO/IEC 11179. Tato norma standardizuje a nabízí možný způsob uchovávání informace o určitém datovém prvku. V dizertační práci rozeberu základní množinu popisných atributů, které tato norma definuje.

Teze 12. Ke každému modelu z architektury P3A (konceptuální, logický a fyzický) lze udržovat datový slovník na odpovídající úrovni abstrakce.

Na základě specifikace jednotlivých modelů uvedu praktické příklady odpovídajícího datového slovníku, který má v každé úrovni svá specifika. EDD konceptuálního modelu používá obecné, platformově nezávislé obsahy popisných atributů. Oproti tomu fyzická implementace entit a jejich atributů je již popsána v jazyce, který je specifický pro zvolenou implementační platformu. V případě této dizertace zvolím opět popis pro DMBS Oracle.

III. Sémantické mapování modelů různých zdrojů

<u>Teze 13.</u> Informační systémy obecně vyžadují design, integraci a údržbu velmi komplexních aplikačních artefaktů. Aby s nimi bylo tyto aktivity možné efektivně provádět, je třeba využít nástroj, který umožní manipulaci s jejich formálním popisem, tedy model (viz. Teze 12.). Tato manipulace často zahrnuje design transformací mezi jednotlivými modely, který vyžaduje nějakou formu explicitního vyjádření resp. mapování.

Budu předpokládat existenci obecného konceptuální modelu v libovolném MMS – Model management systému. Tvorbu a principy takového modelu vysvětlím a doplním ukázkami.

Hlavními cíli správy modelů je zajistit vhodnou podporu pro řízení změn v rámci modelu a, což je pro naši potřebu důležitější, zajistit mapování objektů mezi dvěma odlišnými modely.

S modelem nebo jeho objekty lze provádět celou řadu operací. Tyto obecné algebraické operace jsou použity v jednotlivých dotazech za účelem manipulace s celými modely nebo jejich mapováními. Tyto operace v práci pouze představím, protože pro pochopení principů transformace dat a jejich následné vizualizace, nejsou klíčové.

<u>Teze 14.</u> Jednotlivé modely spolu mohou souviset. Mezi jejich objekty figurují vazby, které lze definovat a popsat pomocí mapování.

V práci použiji Bernsteinovo rozdělení možných typologií mapování:

- Mapování mezi definicí třídy a relačním schématem za účelem vygenerování
- Mapování mezi XML schématy pro řízení překladu zpráv.
- **Mapování mezi zdroji dat a zprostředkovatelským schématem pro řízení integrací heterogenních dat.**

- Mapování mezi databázovým schématem a plánem jeho budoucí implementace (releasem) za účelem řízení migrace dat.
- Mapování mezi entity relationship (ER) modelem a SQL schématem (logickým a fyzickým) pro navigaci v databázi (BERNSTEIN, 2003)

Hodlám se zabývat právě mapováním mezi modely reprezentujícími různé datové zdroje.

Při jeho popisu a vysvětlení využiji výstupy z technického reportu „A vision of management of complex models“ (BERNSTEIN, HALEVY, POTTINGER, 2000).

Budou to následující definice:

- Mapování modelu, které spojuje model M1 s modelem M2, charakterizuje společný kořenový prvek (tzv. *rootElement*). Ten má dvě jednoznačné vlastnosti vztahu, které ukazují na kořenové objekty M1 a M2. V praxi se tyto vlastnosti označují jako *domainRoot* a *rangeRoot* a identifikují modely, které se váží k mapování.
- Každý objekt mapování má vlastnost, označovanou jako *Expr* (výraz), která vyjadřuje potřebnou transformaci mezi objekty v M1 a M2. Informace obsažená v *Expr*, není předem definovaná. Může to být třeba CQL popis transformace (tedy string), SQL, nebo funkce.
- Každé mapování v obecném modelu mapování má dvě vlastnosti vztahu, které se nazývají domain a range. Ty zahrnují všechny objekty z M1 a M2, tj, ty které jsou odkazovány v rámci *Expr*.

<u>Teze 15.</u> Pokud je mapování plně interpretováno může být následně přetransformováno do programu, který automaticky provádí transformaci dat z instance jednoho modelu do instance modelu druhého.

V této oblasti se budu zabývat problematikou interpretace sémantiky mapování.

Poukážu na výhody i nevýhody úplnosti sémantiky mapování:

- Interpretace sémantiky mapování přináší řadu výhod. Pokud je mapování plně interpretováno (vlastnost *Expr* (viz. východisko Teze 14.) obsahuje formuli určitého matematického systému – logiku, algebru, nebo gramatiku), může být následně přetransformováno do programu, který provádí transformaci dat z instance jednoho modelu do instance modelu druhého, a to i automaticky.

- Na druhou stranu je třeba si uvědomit, že využití sémantiky mapování vnáší do datového modelu již konkrétní definice operací, a tím i značnou komplexitu na úkor obecnosti.

Podrobně se budu zabývat metodikou mapování pomocí operace *Match*, tak jak je specifikována v práci "On Matching Schemas Automatically" (RAHM, BERNSTEIN, 2001).

Zápis a výsledky mapování pomocí operace *Match* předvedu na příkladech.

<u>Teze 16.</u> Při vizualizaci mapování je vhodné volit jazyk UML.

Vyhraním se oproti dalším metodám záznamu mapování a jeho vizualizace:

- Mapování definované pomocí přirozeného jazyka, oproti jazyku UML, inklinuje k vágnosti. Samotné pak neumožňuje následné automatické zpracování ani znovupoužití.

- Transformační algoritmy určené pro generování kódu jsou sice vhodné pro strojové zpracování, k jejich pochopení je však potřeba příkladů, nebo hlubšího zkoumání daného algoritmu. Značná zaujatost jedním směrem, neumožňuje

zpětné generování zdrojového elementu ani možnou rekonciliaci (synchronizaci) modelu.

Demonstruji vizualizaci mapování na obecném příkladě, který je možné implementovat pro konkrétní mapování entit a jejich atributů a následně s ním dále pracovat při automatizaci transformací dat mezi dvěma systémy.

IV. Přehled citovaných informačních zdrojů a výběrová bibliografie

1. Aberdeen Group. 2007. *Customer Data Quality:Roadmap for Growth and Profitability* [online]. A White Paper of A Hartle-Hanks Company. June, 2007. [cit. 2007-07-22]. Dostupný z WWW: <http://research.ittoolbox.com/white-papers/pdfViewer.asp?r=http://hosteddocs.ittoolbox.com/Aberdeen_CDQ.PDF>.

2. ADELMAN, Sid; MOSS, Larisa; ABAI, Majid. 2005. Data Strategy. Addison-Wesley Professional IN, June 25, 2005. ISBN 978-0321240996.

3. AKENHURST, D. H.; KENT, S. 2002. A relational approach to defining transformations in a metamodel. In UML 2002 - The Unified Modeling Language. Model Engineering, Languages, Concepts, and Tools. 5th International Conference, Dresden, Germany, September/October 2002, Proceedings [online]. Springer, J.-M. Jezequel, H. Hussmann, and S. Cook, Eds., vol. 2460 of LNCS, 243-258. [cit. 2007-08-10]. Dostupný z WWW: <http://www.cs.kent.ac.uk/projects/kmf/Documents/uml02transf.pdf>.

4. BENYOVSZKY, Štěpán, Ing. 2003. eProvisioning: Synchronizace obsahu informací mezi nesourodými systémy. In ISSS2003 - Konference Internet ve státní správě a samosprávě. Hradec Králové, 23. 3. 2003. [online]. [cit. 2007-07-20]. Dostupný z WWW: <http://www.isss.cz/archiv/2003/download/prezentace/BENYOVSZKY_clarionet.ppt>.

5. BERNSTEIN, Philip A.; HALEVY, Alon Y.; POTTINGER, Rachel, A. 2000. A vision of management of complex models [online]. SIGMOD Record 29(4):55-63 (2000). [cit. 2007-08-10]. Dostupný z FTP: <ftp://ftp.research.microsoft.com/pub/tr/tr-2000-53.pdf>.

6. BERNSTEIN, Philip, A. 2003. Applying Model Management to Classical Meta Data Problems. In 2003 CIDR Conference [online]. (Microsoft Research, One Microsoft Way) [cit. 2007-08-10]. Dostupný z WWW: <http://research.microsoft.com/~philbe/PBERNSTEINCIDR12ext.pdf>.

7. BOHUSLAV, Jiří. 2006. Metody a procesy čištění dat. IT Systems [online]. Příloha Business Intelligence.7-8/2006. Str 14. [cit. 2007-08-10]. Dostupný z WWW: <http://www.systemonline.cz/business-intelligence/metody-a-procesy-cisteni-dat.htm>. ISSN 1802-615X.

8. DAVENPORT, Thomas, H.; COHEN, Don; JACOBSON, Al. 2005. Competing on Analytics. [online]. Babson Executive Education - Working Knowledge Research Report. May, 2005. 4-5 s. [cit. 2007-09-10]. Dostupný z WWW: <http://www.babsonknowledge.org/analytics.pdf>

9. ECKERSON, Wayne. 2004. Data Profiling: A Tool Worth Buying (Really!). DM Review Magazine [online]. June, 2004 Issue [cit. 2007-08-03]. Dostupný z WWW: <http://www.dmreview.com/article_sub.cfm?articleId=1003990>.

10. HAUSMANN, Hendrik, Jan; KENT, Stuart. 2003 Visualizing Model Mappings in UML In Proceedings of the 2003 ACM symposium on Software visualization 2003, San Diego, California June 11 - 13, 2003. SESSION: All things UML [online]. Strana: 169-178. [cit. 2007-08-11]. Dostupný z WWW: <http://wwwcs.uni-paderborn.de/cs/ag-engels/Papers/2003/Softvis03-HAUSMANNKENT.pdf>. ISBN:1-58113-642-0

11. HORRIDGE, Matthew; RECTOR, Allan; STEVENS, Robert; WROE, Chris. 2004. A Practical Guide To Building OWL Ontologies Using The Proétgé-OWL Plugin and CO-ODE Tools [online]. Edition 1.0. The University Of Manchester. August 27, 2004. [cit. 2009-10-01]. Dostupný z WWW: <http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf>.

12. ISO/IEC 11179 - 1:1999. Information technology — Specification and standardization of data elements — Part 1:Framework for the specification andstandardization of data elements [cit. 2007-04-30]. Dostupný z WWW: <http://metadata-standards.org/11179-1/ISO-IEC_11179-1_1999_IS_E.pdf>.

13. KLEMPA, Tomáš. 2006/2007. Opis jazyka OWL pre reprezentáciu ontológií [online]. Slovenská technická univerzita, Fakulta informatiky a informačných technológií, Ústav informatiky a softvérového inžinierstva. Prispevok k prednaske Znalostné systémy. [cit. 2009-15-02]. Dostupný z WWW: <http://www2.fiit.stuba.sk/~kapustik/ZS/Clanky0607/klempa/index.html>.

14. KYJONKA, Vladimír. 2006. Datová kvalita pod lupou. IT Systems [online]. Příloha Business Intelligence.7-8/2006. Str 16. [cit. 2007-08-11]. Dostupný z WWW: <http://www.systemonline.cz/business-intelligence/datova-kvalita-pod-lupou-1.htm>. ISSN 1802-615X.

15. LUJÁN-MORA, Sergio; VASSILIADIS, Panos; TRUJILLO, J. 2004. Data Mapping Diagrams for Data Warehouse Design with UML. In Congrès ER 2004: conceptual modeling (Shanghai, 8-12 November 2004) [online]. [cit. 2007-09-2]. Dostupný z WWW: <http://www.cs.uoi.gr/~pvassil/publications/2004_ER/ER_2004.pdf>

16. MOHANEC, Martin. 2004. Několik poznámek k porozumění objektového paradigmatu. In Objekty 2004 [online]. Ostrava, VSB-TUO, 2004, s. 189-197. [cit. 2007-08-03]. Dostupný z WWW: <http://objekty.pef.czu.cz/2004/sbornik/03_Molhanec.pdf>. ISBN 80-248-0672-X>.

17. MOHANEC, Martin. 2006. KONCEPTUÁLNÍ MODELOVÁNÍ, FORMÁLNÍ ZÁKLADY A ONTOLOGIE [online]. České vysoké učení technické – FEL. Česká republika. 2006. [cit. 2009-24-02]. Dostupný z WWW: <http://formular-ekf.vsb.cz/formulare/F01/tsw/getfile.php?prispevekid=873>.

18. OLSON, Jack. 2002. Data Profiling: The Data Quality Assurance Analyst's Best Tool. DM Direct Newsletter [online]. December 13, 2002 Issue [cit. 2007-08-09]. Dostupný z WWW: <http://www.dmreview.com/article_sub.cfm?articleId=6156>.

19. PAPÍK, Richard. 2001. Competitive Intelligence, informační služby, Internet a informační profese. Ikaros [online]. 2005. Roč. 5, č. 4 [cit. 2007-07-18]. Dostupný z WWW: <http://www.ikaros.cz/node/739>. URN-NBN:cz-ik739. ISSN 1212-5075.

20. RAHM, Erhard, BERNSTEIN, Philip, A. 2001. On Matching Schemas Automatically [online]. Microsoft Research Technical Report MSR-TR-2001-17. February, 2001. [cit. 2007-08-10]. Dostupný z FTP: <ftp://ftp.research.microsoft.com/pub/tr/tr-2001-17.pdf>.

21. RFC 2822. Internet Message Format [online]. 2001 Resnick, P. April 2001 [cit. 2008-04-28]. 51 s. Dostupný z FTP: <ftp://ftp.rfc-editor.org/in-notes/rfc2822.txt>.

22. RUSSOM, Philip. 2007. Unifying the Practices of Data Profiling, Integration, and Quality (dPIQ) [online]. TDWI Monograph Series. October, 2007. [cit. 2007-12-10]. Dostupný z WWW: <http://download.101com.com/pub/tdwi/Files/TDWI_Monograph_DataFlux_Oct2007.pdf>.

23. SATRAPA, Pavel. 2000. Seriál Regulární výrazy. Root.cz [online]. 2000. [cit. 2008-25-02]. Dostupný z WWW: <http://www.root.cz/serialy/regularni-vyrazy/>. ISSN 1212-8309.

24. SVÁTEK, Vojtěch; LABSKÝ, Martin. 2003. Objektové modely a ontologie - podobnosti a rozdíly [online]. Katedra informačního a znalostního inženýrství, Vysoká škola ekonomická v Praze, nám. W. Churchilla 4, 130 67, Praha 3. [cit. 2009-15-02]. Dostupný z WWW: <http://nb.vse.cz/~svatek/obj03fi.pdf>.

25. SVÁTEK, Vojtěch. 2002. „Ontologie a WWW“ in DATAKON 2002, Brno, 19. – 22. 10. 2002, p. 1–35, ISBN 80-210-2958-7.

V.Příloha: Zkušenostní základna k vypracování dizertační práce

PhDr. Ivan Bartoš

Vzdělání:

- složení rigorózní zkoušky na Filozofické fakultě Univerzity Karlovy (5/2005)
  Název rigorózní práce: "*Aplikace protokolu Z39.50 a perspektivy dalšího rozvoje*"

- absolvent magisterského programu Informační studia a knihovnictví na Filozofické fakultě Univerzity Karlovy *(ukončení studia 2004)*

Stipendia:

- stipendium UK *(9/2004 – 12/2004)*
  s příspěvkem Nadace „Nadání Josefa, Marie a Zdeňky Hlávkových"

studijní pobyt: Computer Science Faculty, University of New Orleans- New Orleans, Luisiana, USA *(předměty: Data Models and Database Systems, Introduction to Artificial Intelligence, Computer Security)*

Členství v profesních organizacích:

Z39.50 Implementers Group – Czechia – Národní výzkum aplikace protokolu Z39.50 a souvisejících technologií – asociace implementátorů. "Specialista na standardizaci a implementaci Z3950 – konzultant"

Další odborné aktivity:

Přednášková a publikační činnost s tématikou: *sdílená katalogizace v knihovnách, Z39.50 architektura, protokoly pro získávání, přenos a ukládání informací mezi heterogenními zdroji dat.*

Přednášky pro odbornou komunitu:

"Prezentace - Information Retrieval (Z39.50): Application Service Definition and Protocol Specification" - ZIG CZ (http://www.stk.cz/ZIG/). *(24/10/2002, 2/11/2002)*

"Vybrané aplikace protokolu Z39.50l" - "Úvod do protokolu Z39.50" přednášky pod záštitou Státní technické knihovny v Praze *(2003).*

Semináře vedené v rámci studia UISK - FF UK:

"Datové modely a databázové systémy I",„“ Datové modely a databázové systémy II" - semináře, Ústav informačních studií a knihovnictví, Filozofická fakulta, Karlova univerzita (*2005-2006)*

“Information služby internetu“ - seminář, Ústav informačních studií a knihovnictví, Filozofická fakulta, Karlova univerzita (*2005-2006)*

Publikační činnost:

1. Bartoš, Ivan; Šmilauer, Bohdan Ing. Získávání dat z informačních systémů (Z39.50): Definice aplikačních služeb a specifikace protokolu - volný výklad původní normy. ZIG - CR. 2002, Dostupný z WWW: <http://www.stk.cz/ZIG/Z39.50.zip>

2. Bartoš, Ivan. Aplikace protokolu Z39.50 a perspektivy dalšího rozvoje. 2003, Praha. White paper pro účely ZIG - CR. 150s.

3. Bartoš, Ivan; Šmilauer, Bohdan Ing. Úvod do protokolu Z39.50. In *Moderní informační a komunikační technologie v knihovnictví 2003*. 2003. s. 71-87.

4. Bartoš, Ivan. Information Search and Retrieval in Libraries. Trends, Theory and Practice - Perspectives for Further Development. 2004, New Orleans. LA. USA. White paper, 90 pg.

Praxe v souvisejících oborech:

Datum (od-do)	*2010 (listopad) –*
Název a adresa zaměstnavatele	***T-Mobile Czech Republic a.s. - Tomíčkova 2144/1,149 00 Praha 4, URL:*** *http://www.t-mobile.cz/*
Oblast činnosti firmy	Telekomunikace
Náplň práce/zastávaná pozice	Funkční architekt
Hlavní zodpovědnost	Řešení projektů velkých rozměrů s dopadem na miliony uživatelů zahrnující stovky nezávislých subsystémů a platforem, které zprostředkovávají specifické funkcionality. Řešení jejich vzájemné integrace a interakce včetně analýz pracnosti a hodnocení finančních nabídek k realizaci. (Leading solution designer, Solution analyst a Functional Architect) Senior pozice

Datum (od-do)	*2009 (listopad) – 2010(listopad)*
Název a adresa zaměstnavatele	***MobilKom, a.s. - Corso Karlín, Křižíkova 237/36 a, 186 00 Praha 8 – Karlín, URL:*** *http://www.mobilkom.cz/*
Oblast činnosti firmy	Telekomunikace
Náplň práce/zastávaná pozice	Senior MIS Oracle analytik
Hlavní zodpovědnost	Analytik a architekt oddělení Main Marketing Information System. Projekty orientovány na platformu Oracle (Oracle BI & Answers. Billing) DWH migrace dat ze systémů SAP (ERP, CRM) a networking data. Business a data analýzy, UML návrhy řešení projektů, architektura systémů a loadů. Vývoj. Senior pozice.

Datum (od-do)	*2005 (květen) – 2009 (srpen)*
Název a adresa zaměstnavatele	***MonsterWorldwide - Václavské náměstí 11, 1 10 00 Praha 1, Česká republika - Tel.: +420 2390140481, URL:*** *www.monster.com*
Oblast činnosti firmy	Worldwide on-line recruiter agency
Náplň práce/zastávaná pozice	Senior Oracle Database Engineer
Hlavní zodpovědnost	Strategic Marketing Automation Systems, UNICA Affinium Suite integrace. Design architektury, databází. Oracle database administrace, Oracle Warehouse Builder – OLAP, Oracle BI, ETL procesy, design testovacího frameworku, SQL, T-SQL, PL/SQL vývoj a konzultace, projektově orientované řízení týmů. Senior pozice.

Datum (od-do)	*2003 – 2005 (květen)*
Název a adresa zaměstnavatele	***NEWTON INFORMATION TECHNOLOGY s.r.o. Politických vězňů 10, 1 10 00 Praha 1, Česká republika Tel.: +420222192110, Fax: +420222192192, URL:*** *www.newtonit.cz*
Oblast činnosti firmy	Media monitoring, analýza médií
Náplň práce/zastávaná pozice	Databázový administrátor a designér/ Analytik
Hlavní zodpovědnost	Administrátor a designér databáze zdrojů (včetně front-endového aplikačního software), Senior pozice.

Datum (od-do)	*2004 (září – prosinec)*
Název a adresa zaměstnavatele	***Earl K. Long Library - Lakefront Campus, 2000 Lakeshore Drive, New Orleans, Luisiana70148, USA - Tel.: 1-504 280 6556 Fax: 1-504 280 7277 URL:*** *http://library.uno.edu/*
Oblast činnosti firmy	University of New Orleans library – informační a knihovnické služby
Náplň práce/zastávaná pozice	Tutor, databázový designer, datový analytik
Hlavní zodpovědnost	Projekty: "Electronic Resources Statistics", "Special Collections Registration", design a vývoj databáze + vývoj dílčích aplikací pro prostředí WWW (SQL, ASP, Java), Kompilace statistik a analýza registrací a užívání komerčních databází.

Datum (od-do)	*2001 – 2003*
Název a adresa zaměstnavatele	***Státní technická knihovna – Praha, Mariánské náměstí 5, 110 01 Praha 1, Česká republika - Tel.: +420 221 663 111 Fax: +420 222 221 340 E-mail:*** *techlib@stk.cz* ***URL:*** *www.stk.cz*
Oblast činnosti firmy	Informační a knihovnické služby
Náplň práce/zastávaná pozice	Spolupráce na projektech v rámci LI01018 URL: *http://info.jib.cz/o-projektu/projekt/portal-stm-projekt-li01018*
Hlavní zodpovědnost	Přednášky, výzkum, publikace (SFX, STM Catalogue, MetaLib). viz. publikační činnost.