Zápis z obhajoby disertace

Doktorand:	**PhDr. Ivan Bartoš,** **Ústav informačních studií a knihovnictví, FF UK**
Datum konání obhajoby:	26.9.2012
Místo konání obhajoby:	Filozofická fakulty Univerzity Karlovy, Náměstí J. Palacha, Praha 1.
Téma disertace:	Metodologie a problémy při transformaci dat a určení jejich významu v rámci integrace heterogenních informačních zdrojů
Vedoucí práce:	Doc. PhDr. Richard Papík, Ph.D.
Oponenti:	Dr. Jan Dvořák, Ph.D. Ing. Miroslav Bureš, Ph.D.
Přítomni:	doc. RNDr. Jiří Souček, DrSc. PhDr. Barbora Drobíková, Ph.D. doc. PhDr. Richard Papík, Ph.D. ing. Martin Souček, Ph.D. doc. PhDr. Rudolf Vlasák Dr. Jan Dvořák, Ph.D. Ing. Miroslav Bureš, Ph.D. PhDr. Helena Lipková, Ph.D. a zástupci veřejnosti
Komise:	doc. RNDr. Jiří Souček, DrSc. PhDr. Barbora Drobíková, Ph.D. doc. PhDr. Richard Papík, Ph.D. ing. Martin Souček, Ph.D. doc. PhDr. Rudolf Vlasák
Zapisovatel:	PhDr. Helena Lipková, Ph.D.

1. Obhajobu práce zahájil **přivítáním doktoranda a ostatních přítomných** doc. Souček.

2. Doc. Souček požádal vedoucího práce, doc. R. Papíka, o **představení uchazeče, shrnutí průběhu studia a stručným obsahem jeho disertační práce**

Doc. Papík v základních rysech popsal cíl disertační práce a její základní části. Poté doc. Papík stručně představil doktoranda.

3. Doc. Souček požádal doktoranda **o představení podstatných bodů jeho disertační práce**, a to v cca 5-10 min prezentaci

PhDr.. Bartoš na úvod zdůvodnil výběr tématu – integrace dat nabývá stále více na aktuálnosti při propojování IS.
Cíl: nabídnout komplexní řešení integrace heterogenních databázových zdrojů (end-to-end řešení)
Vymezení: doposud nebylo řešeno – doposud publikované práce se zabývají buď konkrétní implementací nebo dílčí částí celého procesu
Metoda: metodologie ověřená v praxi. Jednotlivé kroky jsou řešeny na konkrétních příkladech. (Experimentální ověření – řešené projekty.)
Obsah přednesené prezentace – viz příloha č. 1 tohoto zápisu.
Doktorand dále uvedl, že práce byla experimentálně ověřena – Newton IT (akvizice několika firem HR) – koncepty v kap 3 a 4 – aplikovány ve společnosti T-Mobile.

4. Doc. Souček požádal **oponenty o stručné přednesení závěrů jejich oponentských posudků**. Jako první bylo slovo uděleno Dr. Janu Dvořákovi.

4. A Oponent Dr. Jan Dvořák, Ph.D.

Dr. Dvořák přednesl své stanovisko k práci, přičemž vyšel ze svého oponentského posudku:
„Předložená disertační práce se zabývá problematikou dat – jejich struktury, významu a kvality – a to zejména ve světle stále častěji se objevující potřeby výměny informací, která vede k datovým transformacím.
Práce je věnována následujícím úlohám:

1. *Klasifikace dat a analýza jejich kvality. Zde se autor věnuje profilaci a validaci dat, a to z teoretické i praktické stránky.*
2. *Transformace dat ze zdrojového do cílového informačního systému. Zde autor probírá všechny obvyklé fáze procesu: extrakci dat, jejich čištění, opravu, konverzi a transformaci do struktury cílového systému.*
3. *Modely dat a datový slovník. Autor prezentuje tři používané úrovně modelování: konceptuální, logickou i fyzickou. Konceptuální úroveň prezentuje též jako oblast aplikace ontologií.*
4. *Sémantické mapování datových modelů. Autor se věnuje problematice manipulace s modely, zejména rozpoznání toho, které části dvou modelů obsahují jednu informaci. Následně tuto informaci využívá při konstrukci pravidel mapování včetně jejich vizualizace.*

Autor zvolil příklad dvou firem a dopadu jejich fúze na jejich informační prostředí. Příklad je srozumitelný, názorný a umožňuje velmi dobře ilustrovat probíraná témata.
Autor skutečně identifikuje hlavní úskalí integrace heterogenních informačních zdrojů: potřebu posoudit kvalitu dat (aby se integrace účastnila jen data dostatečně kvalitní) a potřebu zachovat informaci obsaženou v datech, přestože datová struktura se mění. Předpokladem je mít k dispozici dostatečně vypovídající model dat z hlediska struktury i obsahu.
Autorův výklad je souvislý a zvolené pole dobře pokrývá. Relativní závažnost jednotlivých potenciálních problémů se dobře odráží v míře pozornosti, která je jim v práci věnována. Autor bohužel ne vždy jasně odděluje vlastní přínos od prezentace poznatků a přístupů z odborné literatury.
Autor v práci dává vedle výkladu prostor i ukázkám kódu. Příkladem je balíček PKG_TABLE_ANALYZE v jazyku PL/SQL pro systém řízení báze dat Oracle, který skutečně produkuje relevantní shrnující veličiny pro jednotlivé sloupce databázové tabulky. Autorovi lze doporučit, aby balíček publikoval pod některou volnou licencí, neboť jde o skutečně zdařilý a užitečný nástroj. Zároveň je třeba podotknout, že obvykle se podobně rozsáhlé ukázky kódu umisťují do příloh práce, případně jsou zpřístupněny na Internetu.
***Závěr:** Práce dobře dokumentuje autorovu schopnost teoretické syntézy i jeho praktickou zkušenost s popisovanou oblastí. Přes výše uvedené mírné výhrady ji doporučuji k přijetí."*

Doc. Souček poděkoval Dr. Dvořákovi za jeho posudek. Poté požádal o vyjádření druhého oponenta, ing. Bureše, Ph.D.

4.B Oponent ing. M. Bureš, Ph.D.

Ing. Bureš přečetl svůj oponentský posudek:

„Autor navrhuje možnosti, jak realizovat jednotlivé fáze integrace dat a dokumentuje je na řadě praktických příkladů, které vycházejí z modelové úlohy představené na začátku textu. Práce se strukturou pohybuje mezi rozsáhlou rešerší existujících konceptů a způsobů řešení a návrhem vlastního řešení. Po konzultaci se školitelem toto odpovídá zadání práce.

Mezi klady práce patří vysoká relevance řešené problematiky, přímá vazba na praxi a z toho plynoucí i praktická aplikovatelost a vysoká užitečnost výsledků. Z práce jsou zřejmé praktické zkušenosti autora s řešenou problematikou.

Dalším pozitivem práce je ucelený pohled na celý cyklus datové integrace, pokrývající fáze analýzy kvality dat ve zdrojovém systému, výběru a přenosu klasifikované informace a integrace datových schémat různých zdrojových systémů do sjednocujícího modelu včetně vytváření odpovídajících metadat. Tento cyklus je v práci logicky a přehledně rozdělen do čtyř částí. Takto široce zabrané téma je však zároveň i jistou nástrahou - podle mého názoru je téma pro detailní zpracování na úrovni disertační práce příliš rozsáhlé. Každá ze čtyř částí by pravděpodobně vystačila jako výzkumné téma na samostatnou disertační práci. Z toho se odvíjí i úroveň detailu zpracování jednotlivých částí problematiky. První část týkající se analýzy kvality dat ve zdrojovém systému je zpracována nejrozsáhleji. Další části práce jsou stručnější. Nicméně cílem předložené práce je poskytnout integrující pohled na celou problematiku a to práce splňuje dle zadání.

Velkým záporem práce je bohužel nedostatečná vazba na existující výzkum v dané oblasti. Práce sice obsahuje řadu referencí, uváděné řešení však není vztaženo k předchozímu výzkumu. V oblasti datové integrace existuje velké množství předchozích teoretických prací týkajících se dané tématiky. Z pohledu metodiky výzkumu je nezbytné provést rešerši existujících prací a vymezit se vůči stávajícímu výzkumu. To v práci není explicitně provedeno. Zároveň se struktura práce nedrží standardní konvence pro výzkumnou práci, kterou je rešerše stávajícího stavu, definice řešeného problému, formální popis problematiky, návrh řešení, experimentální ověření navrženého řešení a zhodnocení výsledků. Vzhledem k zadání práce a jejímu cíli je ale pravděpodobně vhodnější struktura, kterou zvolil autor.

Biografii autora uvedenou na konci úvodu bych přesunul do přílohy práce. Pojmy data, informace a znalosti (strana 10) je možné použít v práci ve vlastním významu, ale je zapotřebí definovat jejich význam formálněji, ne pouze pomocí příkladu. Při popisu modelové úlohy (kapitola Vzorový projekt, strana 14) bych doporučil na začátku popsat její význam v práci. Doplňující informace k projektu v další kapitole je možné sloučit s informacemi v kapitole popisujícími modelovou úlohu. V práci je řada kvalitních výstupů dobře aplikovatelných v praxi, které svědčí o autorově zkušenosti s danou problematikou invenci v dané oblasti. V textu práce je použita řada neurčitých formulací, někdy v příkladech, na kterých je problematika dokumentována (např. strany 26, 46 nebo 87). Doporučil bych do elektronických příloh práce uvést konkrétní data modelové úlohy a výstupy příkladů a závěry prezentovat exaktněji. Závěr práce je poměrně stručný a bylo by vhodné jej rozšířit o rozsáhlejší zhodnocení navrženého řešení. Autor uvádí, že prezentovaná metoda byla úspěšně aplikována na několika IT projektech datové integrace. Zde bych v práci očekával komentář týkající se osobní invence autora na těchto projektech a přesné vymezení, že prezentovaná metoda je dílem autora ve vztahu k těmto projektům. Praktickou aplikaci metody je pak možné interpretovat jako její experimentální ověření, je zde však potřeba dodat odpovídající komentář. Co se týče formální úrovně práce, text obsahuje řadu neformálních obratů (příkladem jsou výrazy uvedené v uvozovkách na straně 14 nebo 87), které by bylo vhodné nahradit přesnějšími formulacemi. V uvedených UML diagramech je zcela zbytečné uvádět metadata týkající se diagramu, kdo je autorem, kdy byl diagram vytvořen a další (například strany 22 a 23). Na straně 77 je poškozený a nečitelný obrázek diagramu. Tyto nedostatky jsou v práci zcela zbytečné. Na stranách 31 a 65 jsou dlouhé výpisy zdrojového kódu, které by bylo třeba přemístit do příloh práce. U těchto výpisů je zapotřebí také většího komentáře vysvětlujícího jednotlivé části zdrojového kódu a především jejich vazbu na text práce. V přílohách práce by bylo vhodné uvést i seznam použitých zkratek (přestože jsou při použití v textu vysvětleny v poznámce pod čarou). Zvolené řádkování zbytečně plýtvá prostorem na stránce. Přes výše uvedené nedostatky je práce relevantní jak tématem, tak prezentovanými závěry a je vysoce prakticky použitelná.

Vzhledem k výše uvedenému předloženou disertační práci doporučuji k obhajobě, pokud student u obhajoby předloží dokumentaci nejrelevantnější předchozí práce týkající se výzkumu v dané oblasti a vymezí své závěry a řešení vůči těmto souvisejícím pracím. Dále přesně vymezí, jaký je jeho autorský podíl na vývoji této metody v rámci uvedených projektů, na kterých byla aplikována, jaké části metody jsou jeho invencí a vyjádří se k dalším vyjmenovaným nedostatkům práce uvedených v tomto posudku.“

Ing. M. Bureš na závěr konstatuje, že drtivou většinu jeho odpovědí na své připomínky nalezl již v úvodní prezentaci doktoranda a je se způsobem i obsahem těchto vyjádření spokojen.

Doc. Souček děkuje oponentovi za jeho posudek a dává slovo doktorandovi, aby se vyjádřil k předneseným připomínkám.

5. PhDr. Bartoš děkuje za slovo i oponentské posudky a **vyjadřuje se k základním připomínkám oponentů**:

   A. Vyjádření k posudku dr. Dvořáka:
   - kód se pokusil aplikovat v praxi – právě prochází procesem schválení a dopracovává se. Vše je dostupné jako volná licence.

   B. Vyjádření k posudku ing. Bureše:
   - většina odpovědí obsažena v prezentaci k obhajobě disertační práce (předložena přítomným i v tištěné podobě). K odlišení definice data/ informace: data pojímá v práci v kontextu klasické informační vědy. Informace = data v kontextu (data „obalená“ sémantikou).

6. Doc. Souček děkuje za adresování připomínek oponentů a táže se ing. Bureše, **zda** jsou jeho **připomínky zcela a uspokojivě zodpovězeny.** Ing. Bureš konstatuje, že z předloženého materiálu je zřejmé, že práce přináší nové, inovativní přínosy – n-to-n řešení, které doposud nebylo publikováno. Autorský podíl v posudku nezpochybňoval, pouze ho chtěl více okomentovat – i to bylo v rámci obhajoby splněno. Drobné výtky – není potřeba dále komentovat. Autor zodpověděl výtky předložené v oponentuře.
   Dr. Dvořák také konstatoval, že jeho připomínka byla zodpovězena.

7. Doc. Souček **otevírá všeobecnou diskusi.**
   Doc. Vlasák vznáší dotaz týkající se definičního pole : data, informace, znalosti: a jeho využití v rámci práce. Doktorand mimo jiné uvádí, že nastavení těchto definic pomáhá k přesnému a úplnému popisu IS a z tohoto důvodu byly tyto definice zakomponovány do celého rámce předloženého textu. Tento přístup je důležitý i z hlediska pochopení několikavrstvého modelování systémů.

   Doc. Papík navazuje otázkou na využití znalostních systémů ve společnosti T-Mobile. Doktorand odpovídá, že v rámci společnosti mají systémy s plochou strukturou, ale se zakomponovanou funkcí business inteligence, v rámci které se predikuje např. business model na příští rok. Nad doplňkový dotaz. doc. Papíka, zda jsou reporty vytvářeny strojově nebo s lidskou pomocí doktorand odpovídá, že reporty jsou definovány uživatelem. Z těchto reportů jsou získávány znalosti - tam je nutný kognitivní zásah člověka – zde už není systém automatický (firemní politika, odd. marketingu atd.).

   Doc. Souček doporučuje držet se v diskusi více tematiky práce, týkající se datových modelů. Komentář doc. Součka k práci se týká složitosti úlohy datového a konceptuálního modelování systémů s ohledem na pozdější možnou integraci systémů. Význam: 1. navrhování systémů, které v budou v budoucnosti spojovány (=všechny IS), 2. při sjednocování systémů dosavadních. Dotaz: uvažuje nad určitou metodologií pro tvorbu IS, aby byly sjednotitelé,až tento problém vznikne? Doktorand ve své odpovědi odkazuje na existující ISO normy, příp. číselníky v případě knihoven (autority) atd.

   Shrnutí obhajoby:
   Ing.Bureš: práce představuje dobrý teoretický základ pro toho, kdo chce navrhovat systém. Přínosem práce je zejména to, že se jedná o otevřenou metodiku s širokou využitelností.
   Doc. Souček: práce se pohybuje na rozhraní mezi čistou vědou (ve které jsou všechny práce vždy publikovány) a firemními technologiemi (částečně publikovány). Konstatuje složitost hodnocení

původnosti v této situaci, zároveň však doporučuje, aby se firemní literatura brala jako zcela jiná oblast, protože prvky z firemní literatury proniknou do vědecké sféry až když jsou plně ve vědecké práci publikovány (viz dřívější zkušenost s např. asymetrickými kryptografickými systémy - do doby než byly publikovány ve vědeckém světě nebyly známy, byly pouze „utajeny“ ve firemním světě).
Dle názoru oponentů byly zodpovězeny všechny připomínky, vyhovujícím způsobem byly zodpovězeny i dotazy z pléna.

Doc. Souček ukončuje veřejnou část zasedání.
Nečlenové komise opouštějí místnost.

8. **Tajné hlasování členů komise.**

9. Doc. Souček **seznamuje veřejnost s hlasováním komise**:

10. Doc. Souček **ukončuje obhajobu disertační práce** PhDr Ivana Bartoše a blahopřeje mu k úspěšnému absolutoriu.Komise navrhla udělit titul doktor (Ph.D.).

Podpis předsedy komise: