

# Dimense a atributy kvality dat

Jaroslav Král

Katedra SW inženýrství MFF UK

# Co je kvalitita, ISO 8402

- Characteristics of an entity as a whole that give the capability to satisfy explicit and implicit needs:
  - Quality of an entity is a subjective concept dependent on requirements that the user of the entity requests in an implicit or explicit manner.
  - Quality is a multidimensional concept tied to various characteristics.

# Co je kvalita

- Již zde cítíme jistý rozpor.
  - Kvalita dat může být příliš vázána na jednu aplikaci (ale někdy jinak nelze – pojišťovny potřebují větší soubory)
  - Chceme, nějaký atribut kvality (charakteristika, dimenze) byl pokud možno použitelný obecně (jako např. rozptyl)
- O tomto problému se vedou ostré diskuse, viz např. projekt SQuaRE v ISO (ISO 2500xx jako náhrada ISO 9126)

# Atributy kvality dat

Snažíme se o takové atributy kvality, které

- mají význam pokud možno pro řadu různých aplikací pracujících s danými daty,
- nejsou pokud možno přímo vázány na potřeby jediné určité konkrétní aplikace (nejsou vstupem či parametrem algoritmů této aplikace).

# Hlavní zdroje

- [mitiq.mit.edu](http://mitiq.mit.edu), hlavní pracoviště na MIT
- [www.Data.QualityAct.US](http://www.Data.QualityAct.US),
  - US zákon. Stanovuje závazná pravidla pro měření kvality dat ve státní správě
- [www.iqconference.org](http://www.iqconference.org)
- Leo Pipino, Yang W. Lee, Richard Y. Wang: **Data quality assessment.**  
*Communications of the ACM* 45(4): 211-218 (2002), lze získat přes portál ACM

# Nejaktivnější pracoviště: MIT

- **Information Quality at MIT**
  - MIT Information Quality (MITIQ) Program
  - MIT Total Data Quality Management Program
  - IQ Conferences

# Hlavní problémy

- Není jasný rozdíl mezi data quality a information quality. My proto budeme mluvit i o kvalitě informací
  - Tendence chápat jako rozdílné problémy. To může být důležité pro sémantický web.
- Není dost zkušeností, co za míru kvality dat a informací považovat, jak to měřit a jak používat (např. při spojování dat z různých zdrojů)
- Kdy je potřeba stanovit míry kvality legislativně,
  - Nekvalitní data mohou vést ke ztrátám, i životů

# Výzkum kvality dat a informací

Příklady sekcí z IQConference  
2004, koná se každý rok, letos po  
desáté

# Sekce na IQ Conference 2004

- **WEB/INTERNET QUALITY**
- *Session Chair: Craig Fisher, Marist College*
- **Website Quality Assessment Criteria**
- **Analyzing Information Quality In Virtual Service Networks With Qualitative Interview Data**
- **Web Design vs. Web Quality**

# Sekce na IQ Conference 2004

## **IQ RESEARCH FRONTIER**

*Session Chair: Jennifer Long, University of Toronto*

**Simulations of the Relationship Between an Information System's Input Accuracy and Its Output Accuracy**

**Metadata Quality For Federated Collections**

**Logical Interdependence of Some Attributes of Data/Information Quality**

# Sekce na IQ Conference 2004

## **COST BASED CASES**

Session Chair: Latif Hakim, University of  
Southern Queensland, Australia

Beyond Business Process Reengineering

Using the Data Quality Scorecard as a  
Negotiation Strategy

Data Mining, Dirty Data, and Costs

# Co lze vysledovat

- Kvalita dat a informací je drahá záležitost
- Nutnost zohledňovat architekturu systému
- Samozřejmě úzká vazba mezi kvalitou dat a informací
- Při pohledu na celý program konference je nápadný poměrně malý počet výsledků pro servisně orientované systémy (a tedy možná i pro sémantický web).

# Proč se problém kvality dat (a informací) stává rozhodující až nyní

- Bez vyřešení problému, jak data ukládat, vyhledávat a prezentovat, nemělo dříve řešení otázky kvality dat smysl.
- Prvé aplikace databází se převážně týkaly operativy, jako je účetnictví nebo skladové hospodářství. Tam bylo z podstaty věci a zavedenými postupy zajištěno, že data musela být správná – kvalitní, jinak byla nepoužitelná.
- U nás jsme to trochu podcenili

# Data se uplatňují ve státní správě i managementu

- Je nutné zajistit nejen ochranu dat, ale také jejich kvalitu a zavést procedury jak jednat, není-li kvalita dat ideální ale musí se používat
- Dat je mnoho a jsou na webu nutně všelijaká, přesto ale nejsou bezcenná
  - Někdy obsahuje krátký drb více informace než dlouhatánská zpráva

# Proč se problémem kvality dat stává rozhodující 2

- Nebyl dostatečně rozvinut pojmový aparát umožňující specifikovat různé aspekty a dimenze kvality dat.
  - Jak uvidíme, není v tomto směru dnes, přes značný pokrok, dosud dostatečně jasno a je nutný další výzkum a také hodnocení praktických případů zaměřený na kvalitu informací záviselých na daných datech.

# Proč se problém kvality dat stává rozhodující 3

- Chyběly
  - metody a způsoby zápisu atributů kvality dat do metadat (např. RDF) a
  - vědomí důsledků statistických vlastností a jiných metrik kvality datových souborů pro aplikace využívající data určité kvality

# Kvalita dat, věcné problémy

- V managementu se *musí* používat data, která nejsou zcela spolehlivá a relevantní
- Podpora managementu se stává hlavním úkolem informatiky.
- Ukládání a využívání dat operativy je už do značné míry vyřešeným úkolem (neplatí pro zábavu a zčásti i web)

# Formáty metrik

- **Příslušnost ke třídě** (například výskyt určitého znaku, třeba čísla tramvaje)
- **Fuzzy** (dobrý, lepší, nejlepší) – prvek uspořádané množiny, pro níž je jedinou přípustnou operací operace porovnání.
- **Intervalové** (například teplota).
- **Číselné** – jsou povoleny všechny aritmetické operace. Příkladem je rozsah souboru nebo jeho průměrná hodnota.

*Metriky kvality dat jsou většinou fuzzy nebo číselné. Fuzzy metriky jsou subjektivní, jsou získána kvalifikovaným odhadem experta*

# Řízení kvality dat (informací)

- Rozhodnutí o metrikách a procesech jejich měření (assessment) a nápravných opatřeních (control)
- Sběr a čištění (zlepšování jejich kvality - data cleansing)
- Odvozené procesy pro informace založené na zpracování daných dat
- Rozhodnutí o modernizaci nebo zrušení používaných metrik a postupů jejich měření-získávání

# Obor se rychle vyvíjí

- Není shoda o tom, jak metriky třídít

# Subjektivní a objektivní metriky

1. **Objektivní metriky** jsou metriky které lze vždy znovu vypočítat z dat, kterých se týkají.
  - Jsou to často statistické charakteristiky datového souboru (rozsah souboru, průměr, rozptyl, výběrové momenty, např.  $\sum_i x_i^3$ , korelace, atd.). Objektivní metriky jsou obvykle číselné.
2. Objektivní metriky kvality dat odpovídají **externím metrikám** kvality softwaru ve smyslu ISO 9126-1 (např. délka programu)

# Subjektivní a objektivní metriky

**Subjektivní metriky** jsou (fuzzy) metriky získané od hodnotitelů, jsou to např. metriky hodnotící způsob, jakým data vznikla, případně kvalitu zdroje dat. Subjektivní jsou i metriky hodnotící důvěryhodnost dat, stupeň jejich utajení, dostupnost, atd.

Subjektivní metriky odpovídají **metrikám interním** (in process metrics, např. doba řešení, pracnost) podle ISO 9126

# Subjektivní a objektivní metriky

Hranice mezi subjektivními a objektivními metrikami není striktní.

- Pokud máme dostatečně rozsáhlý soubor, můžeme jeho střední hodnotu a směrodatnou odchylku vypočítat a uložit do metadat.
- V opačném případě musíme použít kvalifikovaný odhad, tj. postupovat jako v případě subjektivních metrik. Fakt, že se takto postupovalo, by měl být zaznamenán

# Subjektivní metriky

- Přívlastek *„subjektivní“* má v případě metrik kvality dat jisté oprávnění, poněvadž tyto metriky většinou nevznikají měřením prostřednictvím nějakého technického procesu, ale je de facto subjektivním hodnocením vlastností dat experty založenou na zkušenostech a nikoliv na měření v běžném slova smyslu.
- Pro zkvalitnění dat je i v tomto případě nutno specifikovat proces „měření“, mnohdy zákonem. Často s použitím komplikovaných dotazníků

# Objektivní metriky, data cleansing

- Mezi objektivní metriky patří takové vlastnosti, které lze vypočítat z dat samotných. Tyto metriky se často používají při zlepšování kvality dat.
- „Zlepšováním“ či čišťením dat (data cleansing/cleaning) se míní takové operace jako odstraňování okrajových dat, doplňování dat do časových řad, atd.

# Měření subjektivních metrik, quality assessment

Proces zjišťování subjektivních metrik je nutno standardizovat. To je většinou zajišťováno předpisy (mnohdy na úrovni zákona), které specifikují atributy (dimenze) kvality dat, a postupy, které je nutno při sběru dat a při jejich „čištění“ dodržovat. Příkladem je NRS State Data Quality Standards Checklist

<http://www.doe.mass.edu/acls/smartt/NRSChecklist.pdf>

# Čištění dat

- *Okrajová data* (chyby měření). Jde o postup, kdy se ze souboru vylučují data, která jsou zjevně nesprávná: úmyslně změněná, chybně zanesená (překlepy).
- *Chybějící data*. V tomto případě se do souboru doplní chybějící data, aby bylo možno soubor rozumně zobrazovat (například časové řady nebo matice měření) a přitom nedošlo k “chybným” výsledkům (k významným změnám charakteristik daného souboru).
- *Vyloučení duplicitních dat*
- *Sjednocení formátů*
- *A další*

# Operace nad daty

- *Parciální replikace.* Pokud se data používají pouze pro statistické analýzy (a to je při podpoře managementu obvyklé), lze často soubory dat replikovat pouze částečně (aniž dojde k závažnější chybě). Úspory mohou být dramatické.
- Sjednocování hodnot metrik obecně. Je to vážný problém pro databáze a jde o velmi podceňovaný problém u sémantického webu
- *Existuje na to poměrně rozvinutá teorie a postupy, které se používají především při dolování dat. Problém je, že nevíme, co je nejlepší. Musíme čekat na zkušenosti*

# Nejčastěji používané atributy kvality dat

*Relevantnost* (Relevance) – míra, do jaké míry data splňují účel, pro který jsou používána.

*Přesnost* (Accuracy) – jak přesná jsou používaná data (např. směrodatná odchylka). Kupodivu se neuvažují posunutá data

*Včasnost* (Timeliness) – za jakou dobu lze data aktualizovat.

# Nejčastěji používané atributy kvality dat

*Dostupnost* (Accessibility) – jak jsou již existující data dostupná.

*Porovnatelnost* (Comparability) – metrika hodnotící možnost porovnávat, ale také spojovat data z různých zdrojů.

*Koherence* (Coherence) – metrika vyjadřuje, do jaké míry byla data vytvořena podle z hlediska výsledku stejných pravidel.

*Úplnost* (Completeness) – metrika udávající jaká část potenciálních dat je zachycena v databázi,

# Další metriky

Pro účely statistik, např. FAO, se specifikují další metriky, např. relevance se odvozuje od počtu pozitivních ohlasů, počtu odkazů v publikacích a hodnocení (rate) dostupných statistik

- Kromě výše uvedených metrik se často vyhodnocují další metriky z následující tabulky.

# Kvalita dat, hlavně v e-governmentu

<b>Kategorie</b>	<b>Dimenze</b>
Vnitřní, intrinsická (Intrinsic)	Přesnost (Accuracy) Objektivnost (Objectivity) Důvěryhodnost (Believability) Reputace (Reputation)
Dostupnost (Accessibility)	Dostupnost (Accessibility, též Availability) Bezpečnost přístupu (Access security)
Kontextuální (Contextual)	Relevantnost (Relevancy) Přínos (Value added) Včasnost (Timeliness) Úplnost (Completeness) Rozsah (Amount of data)
Reprezentační (Representational)	Interoperabilita (Interoperability) Srozumitelnost (Easy of Understanding) Výstižná a stručná reprezentace (Concise representation) Konsistentní reprezentace (Consistent representation)

# Problémy s kvalitou dat při dolování dat a na webu

- Jak stanovovat míry kvality, dat jestliže
  - Daná míra má pro různé zdroje různé hodnoty
  - Daná míra je i různě vyhodnocována (jiné procedury vyhodnocování)
  - Daná míra se na některých zdrojích vůbec nevyhodnocuje
  - Je pro nás lepší soubor s milionem údajů a rozptylem 2 nebo soubor s 100 údaji a rozptylem 1? Má smysl tyto soubory spojit?

# Problémy s kvalitou dat při dolování dat a na webu

- Tento problém se různě řeší v datových skladech. Na webu asi závisí na tom, s čím se smíříme a jaké zkušenosti získáme
- Asi budeme muset často rezignovat na požadavek, aby zdroje byly transparentní (nemuseli jsme se o ně zajímat)

# Kvalita informací

- Někdy se ztotožňuje s kvalitou dat
- Není na to jednotný názor. Převažuje názor, že se má kvalita informací chápat jako samostatný problém, který není totožný s kvalitou dat i když s ním úzce souvisí

# Kvalita informací

- Objevuje se tendence k chápání informací jako produktů s dobou života (podobně jako SW systém)
  - Vize, proč se sbírá a vyhodnocuje
  - Konkretizace funkcí a vlastností,
  - Implementace procesů a funkcí pro hodnocení a řízení kvality informací
  - Používání včetně sledování kvality, vylepšování a modifikace
  - Zrušení nebo reinženýring

# Kvalita dat a formulace požadavků na informační systemy

- To, co můžeme uskutečnit je limitováno kvalitou dat více, než jsme ochotni připustit.
- Příklady:
  - Prostředky pro řízení projektů
  - Řízení výrobních procesů

# Kvalita dat a formulace požadavků

- Metoda kritické cesty, viz MSProject, často nefunguje, důvody:
  - Řešitelé podprojektů zadávají doby řešení podle pravidla „to už by muselo být hodně smůly, abych to nestihl“ (nikdo nezadá medián, to by v polovině případů nestihl a byly by postihy)
  - Zadává tedy horní hranici konfidenčního intervalu. Čili se zvažuje přesnost dat.
- Přesto se projekt obvykle nestihne

# Důvody skluzů

- Řešitelé následující etapy nemohou začít řešit úkol dříve (skončí-li předchůdce dříve), než bylo plánováno, neboť musí dokončit jiné úkoly.
- Řešitelé navazující etapy mají na řešení více času než čekali, a proto se zpočátku úkolu příliš nevěnují (efekt líného studenta).

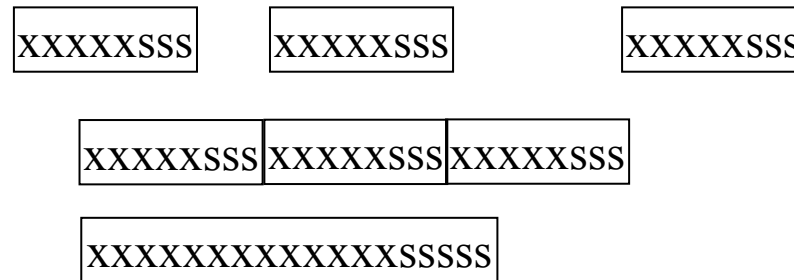
# Důvody skluzů

Většina projektů je omezována nějakým zdrojem Z, o který soutěží více etap. Případné špičky zátěže zdroje Z se řeší tak, že zdroj pracuje střídavě na více úkolech současně (všichni vedoucí etap, kteří Z potřebují, chtějí, aby už proboha Z začal pracovat na jejich úkolu). Výsledkem je, že se řešení všech projektů opozdí (efekt multitaskingu).

# Důvody skluzů

- Je-li někdo hotov dříve, zatluče to, protože hrozí, že příště mu vnutí kratší doby řešení (efekt zpevnování norem)
- Nestihne-li, nedá se nic dělat

# Kritický řetězec – doba řešení je součtem nezávislých n.v.



- Doba řešení  $T$  klasické metody kritické cesty
- $T \geq \Sigma (t_i + 3 \sigma_i)$
- Doba řešení pro kritický řetězec bude většinou

$$\bullet T \approx \Sigma (t_i + 3 \sqrt{\Sigma \sigma_i^2})$$

# PřípojnÉ buffery

A1	A2	A3	A4	A5
----	----	----	----	----

A1	A2	A3	A4	A5	Nárazník projektu
----	----	----	----	----	-------------------

$$\underline{\text{Délka nárazníku}} \approx \Sigma (3\sqrt{\Sigma\sigma_i^2})$$

# Řešení

- Formálněji (obr. 1) můžeme volit
- $NP = \sqrt{(R1^2 + R2^2 + \dots + Rk^2)},$
- kde  $R_i$  jsou rezervy jednotlivých činností na kritické cestě.

# Řešení, kritický řetězec

- Kritický řetěz funguje jen když jsou řešitelé ochotni odhalit rezervy a neskrývat, že jsou hotovi dříve, než se plánovalo.
- Odměny za dodržení termínu (za zkrácení není další bonus, vedlo by to opět k licitování)
- Termín se odvozuje z očekávané doby řešení
- Navíc má každá činnost odhad doby, když jdou věci špatně (horní hranice konfidenčního intervalu)
- Termín pro celý projekt se určuje jako odhad horní hranice konfidenčního intervalu jeho řešení

# Řešení, kritický řetězec

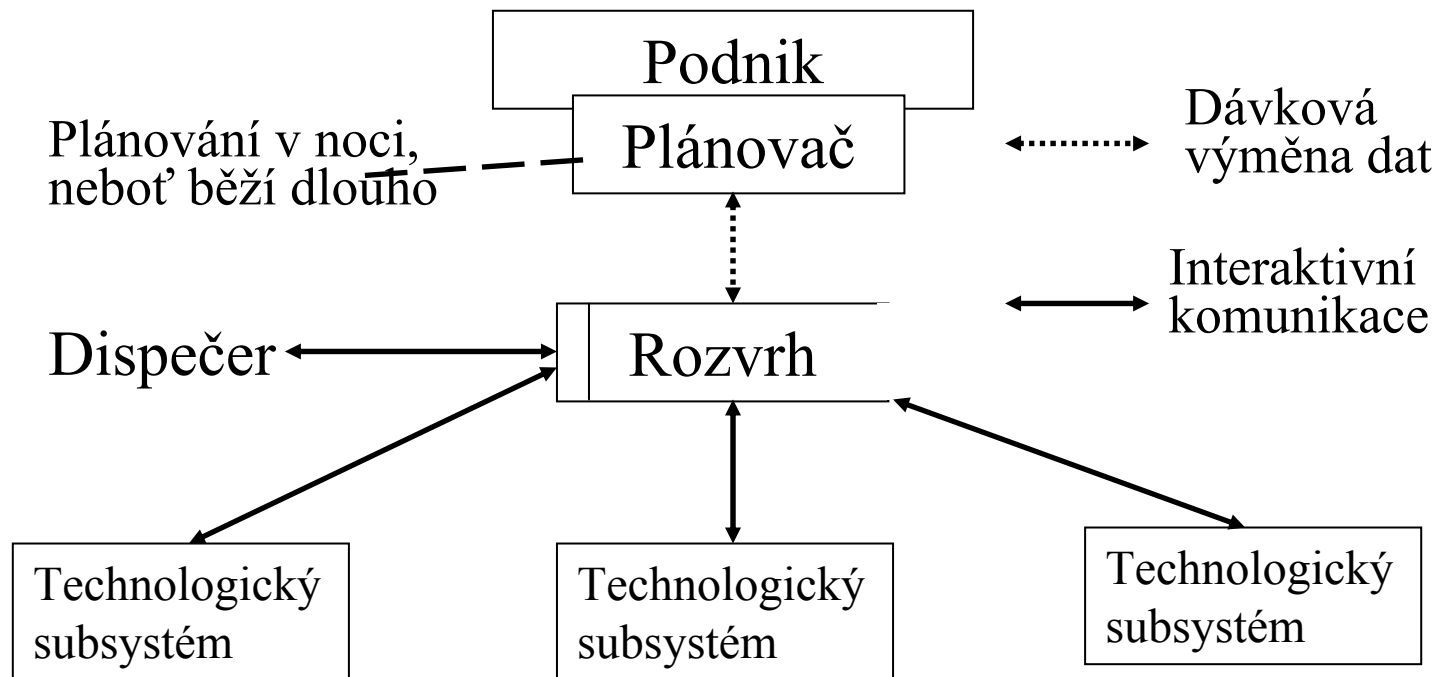
- Každý začne pracovat hned, jak je jeho předchůdce hotov. To ale znamená, že je nutné nějak vyloučit efekt multitaskingu, To se řeší tak, že se dané činnosti postupně stále přesněji oznamuje, kdy bude třeba začít pracovat na daném projektu (činnost je tedy spravována jako autonomní služba). (Zpřesňování dat)
- Práce se odevzdává v okamžiku, kdy je hotova. Její začátek se ale postupně zpřesňuje
- *Důsledek: Dosti často se daří, aby práce trvala přibližně tak dlouho, jako kdyby byly její kroky zcela nezávislé*

## Problém 2: Chybějící a pomalu aktualizovaná data ve výrobě

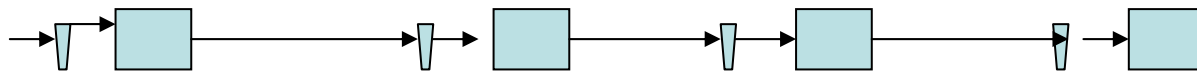
- Rozvrhování na úrovni podniku nemůže mít všechna data a ta co má nemají dobrou kvalitu
  - Cena sběru, některá data se nesbírají
  - Jsou nedostupná (taktilní, zkušenostní)
  - Nedostatečně přesná
- Rozvrhování musí být pomalé (pracnost, interakce s lidmi) a nepřesné v důsledku nekvality dat – je nutné použít úložiště dat (to je navíc potřeba i pokud chceme dát managementu možnost uplatnit své znalosti a intuici při řízení)

# Neinteraktivní komunikace

Spolupráce plánovacích algoritmů s provozem vyžaduje inteligenci při přenosu požadavků



# Řízení na buffery



Mistr sleduje buffery a hledá pro práci, když se bufer nepříjemně zkracuje (řízení na průšvih)

Nutné pro výroby s krátkými seriemi a velkou variabilitou

Segment  
výr. postupu

<b>Prac</b>	P1.z-1 s1 F.j	P1.z s2 D.i-1	P1.z+1 s3	Segment fronty prací na Prac1
<b>Prac</b>	P2.y-1 s4 E.k	P2.y s D.i	P1.z+1 s5	Segment fronty prací na Prac2
<b>Prac</b>	P3.x-1 s6	P3.x s7 D.i+1	P3.x+1 s8	Segment fronty prací na Prac3
Data aktuální výrobní operace D.i				

# Pozorování

- Některé akce musí být dávkové (trvají příliš dlouho).
  - Složitost algoritmů,
  - Nutnost spoluúčasti lidí nebo procesů reálného světa
- Není jasné, zda chápeme důsledky toho, že se jedná o procesy zasahující do reálného světa

# Pozorování

V našem příkladě jsou třeba zásahy dispečera především v těchto případech

- Nečekané/vzácné události - nevyplatí se je zahrnovat do rozvrhování (Vonásek je lempl, Pepa se včera ztřískal, dodavatel to nestihl)
- Kvalita dat
  - Nedostupná, neznámá, nepřesná (mají velký rozptyl)
  - Zřídka potřebná (nevyplatí se sbírat)
  - Pomalu aktualizovatelná (nevčasná)
- Potřeba využít inteligenci lidí jako součásti procesů

# Problémy s kvalitou dat pro řízení

- Relevantnost a včasnost závisí na frekvenci zjišťování nebo na tom, jak je časově náročné data vytvořit (např. data rozvrhu)
- Kvalita dat může implikovat vytvoření datového úložiště v SOA, aby management mohl ovlivňovat chod systému
- ?? Zohledňuje to UML a holistický model v sémantickém webu?

# Problémy s kvalitou dat v státní správě

- Duplicity
- Ochrana dat a zhoršení dostupnosti
- Žádná legislativní ochrana kvality dat
- Katastrofální zhoršení efektů IT ve státní správě
  - Horší obrana proti terorizmu, krizové řízení
  - Obtížné dotazy na efekty zákonů (kvalita škol)
  - Zbytečná buzerace občanů

# Hodnocení kvality škol

- Jak závisí kariéra, příjem, nezaměstnanost na
  - Rozsahu matematiky, typu vzdělání
  - Typu školy, konkrétní škole

Vše by se dalo analyzovat, data existují , ale jsou nedostupná

# Problémy s kvalitou dat pro řízení

- Kvalita dat může implikovat filosofii řešení
  - Kritická cesta a kritický řetězec
- Kvalitu je nutno měřit či odhadovat
- Kvalitu dat můžeme zlepšovat
  - Okrajová data
  - Chybějící data pro parametry, pro regresi
  - Opakovaná data
  - Rozsah dat

# Závěry

- Metriky kvality je žádoucí až nezbytné zahrnout do metadat
- Není zatím jasné, jak při dolování dat a agregátních charakteristikách postupovat při hodnocení kvality souborů dat proměnnou kvalitou. To je zvláště kritické u sémantického webu

# Závěry 2

- Kvalita dat se stává klíčovou částí návrhu IS a architektury SW systémů. Může např. znamenat částečný návrat k datovým úložištím.
- Znamená i změnu filosofie. Antivzor návrhu Ostrov automatizace se stává vzorem.
  - A co UML
- Může podstatně ovlivnit použitelnost sémantického webu a holistických přístupů.
- Srozumitelnost a deklarativnost dat na rozhraních komponent je klíčovou podmínkou použitelnosti business procesů.
  - A co pak objektová orientace?

# Závěr 3

Pokud můžeme soudit, je využití metrik kvality dat a informací zatím v dosti zárodečném stavu i přes poměrně dlouhodobý výzkum

U nás se kvalita dat nepovažuje za důležitou a legislativně se neřeší. To blokuje využití informatiky a ohrožuje informatiku jako celek