Kvalita dat

Výzkum kvality dat

- Kvalita dat je rozsáhlým oborem, který je i tématem legislativní činnosti
- My se zaměříme pouze na klíčové problémy kvality dats přímou vazbou na problematiku informačních systémů
- Kvalita dat a následně kvalita informací se s rozvojem Internetu a podporou manažerských informací stává stále významnější

Co je kvalita, 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.

ISO 9000:2005, PLAIN ENGLISH DICTIONARY

- **The *quality* of something can be determined by comparing a set of inherent characteristics with a set of requirements. If those inherent characteristics meet all requirements, high or excellent quality is achieved. If those characteristics do not meet all requirements, a low or poor level of quality is achieved.**
- ***Quality* is, therefore, a question of degree. As a result, the central quality question is: How well does this set of inherent characteristics comply with this set of requirements?**
  In short, the *quality* of something depends on a set of inherent
  characteristics and a set of requirements and how well the former complies with the latter.
- **According to this definition, *quality* is a relative concept. By linking quality to requirements, ISO 9000 argues that the *quality* of something cannot be established in a vacuum. *Quality* is always *relative to* a set of requirements.**

Co je kvalita

- Již zde cítíme jistý rozpor.
  - Kvalita dat může být příliš vázána na jednu aplikaci a může mít v dané formě smysl jen pro určité uživatele
  - Některé atributy kvality vyžadují velké soubory (př. odhady pro pojistku)
  - Chceme, nějaký atribut kvality (charakteristika, dimense) byl 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

Snažíme se o takové atributy kvality, které

- mají význam pokud možno pro řadu různých aplikací pracujících s danými daty,
- Jsou relevantní či zajímavé pro mnohé uživatele (všechny)
- nejsou pokud možno přímo vázány na potřeby určité konkrétní aplikace

Hlavní zdroje

- mitiq.mit.edu, hlavní pracoviště na MIT
- www.Data.QualityAct.US,
  - US zákon. Stanovuje závazná pravidla pro měření kvality dat ve státní správě
- 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 o kvalitě dat i o kvalitě informací
  - Tendence chápat kvalitu dat a kvalitu informací jako rozdílné problémy.
- Není dost zkušeností, co za míru kvality dat a informací považovat, jak to měřit a jak používat (spojování dat z různých zdrojů)
- Nutnost stanovit míry kvality legislativně, nekvalitní data mohou vést ke ztrátám, i životů
- Umožnit individuální vyhodnocování informací tak, aby vyhodnocující mohl dosáhnout kvality informací optimální právě pro něho
- Kvalita informací závisí na kvalitě dat, z nichž se zjišťuje a na vlastnostech procesů (aplikací), které ji vyhodnocují

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 systém, který ji vyhodnocuje
- Samozřejmě je 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 teď

- 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 v managementu

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

Proč se problém 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 závisejících 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 a mohou býti jinak málo kvalitní
- Podpora managementu se stává hlavním úkolem informatiky.
- Ukládání a využívání dat operativy je už do značné míry vyřešeným úkolem (neplatí pro zábavu a 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, má to být uspořádaná množina

Řízení kvality dat (informací)

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

Obor se rychle vyvíjí

- Není shoda o tom, jak metriky třídit

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 metriky hodnotící způsob, jakým data vznikla, případně kvalitu zdroje dat. Subjektivní jsou 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 založeným 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 cleaning

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

Měření subjektivních metrik, quality assesment

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), nesprávného formátu.
- *Chybějící data*. V tomto případě se do souboru doplní chybějící data, aby bylo možno soubor rozumně zobrazovat (například časové řady) a přitom nedošlo k chybným výsledkům (k významným změnám charakteristik daného souboru). Někdy se doplňují data v řecko-latinských čtvercích
- *Vyloučení duplicitních dat*
- *Sjednocení formátů*
- *A další*

Vyloučení duplicitních dat Sjednocení formátů

- Vylučování duplicitních dat je při nejednotnosti formátů velmi komplikované
  - Např. se dlouho nepodařilo díky tomuto problému vytvořit registr občanů
  - Finanční instituce rozesílají duplicitně dopisy svým klientům o všeobecných službách, které poskytují (ztráty: mnohamilionové výdaje za poštovné, naštvanost klientů)

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 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, týkají se daného problému.

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

Včasnost (Timeliness) – za jakou dobu lze data aktualizovat, jak jsou data aktuální.

Nejčastěji používané atributy kvality dat

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

Obtížný problém díky nesmyslných předpisů pro ochranu privátních dat

Porovnatelnost (Comparablity) – 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 kompatibilních pravidel

Úplnost (Completeness) – metrika udávající jaká část potenciálních dat je zachycena v databázi, případně, zda výběr dat pokrývá „rovnoměrně“ celý výběrový prostor

Další metriky

Pro účely statistik, např. FAO, se specifikují další metriky, např. relevance se odvozuje od počtu positivních 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

Kategorie	**Dimenze**
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 řeší v matematické statistice a různě se 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
  - Spíše metriky výstupů procesů nad daty

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
  - Používání včetně sledování kvality, vylepšování a modifikace
  - Zrušení nebo reinženýring

Dimense kvality informací

Quality Categories	Information Quality Dimensions
Intrinsic IQ	Accuracy, objectivity, believability, reputation
Contextual IQ	Relevancy, value-added, timeliness, completeness, amount of information
Representational IQ	Interpretability, ease of understanding, concise representation, consistency
Accessibility IQ	Access, security

Table 2 Information Quality Categories and Dimensions (Source: Wang, Strong [10])

Dimense kvality informací

Sequence of examination				Examples of attributes of data/information quality		
				Direct Attributes		*Indirect* Attributes
	▼	*Primary* Attributes	*Mandatory* Attributes	**Interpretable**		Legible, user trained, untrained, educated
				Significantly relevant		Concise, current, admissible, secure, appropriate amount
				Critically timely		Obtainable, accessible, style and mode of decision making, individual or collective
			Desirable Attributes	**Critically credible**		Believable, reputable, decision maker's traits: risk averse, passive, hesitant, cautious, prudent, motivated, jumpy
				Acceptably complete (for totality of factors)		
	Irrelevant	*Secondary* Attr.		Economically	**Timely**	Frequency, how much in advance
					Unbiased	Sampling, observation points,
					Accurate (error-free)	Mapping (complete, unambiguous, meaningful, and correct), granularity, age
					Precise	Number of significant digits, dots
					Easy to use	How summarized, detail, text, graph, diagram, picture, media, clarity, order, consistent, homogene, understandable, natural, efficiently encoded

Table 4. Example of a hierarchical result-oriented taxonomy of data or information quality attributes in economical sequence of their examination

Problémy s kvalitou dat pro řízeni

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?

Problémy s kvalitou dat pro řízeni

- 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

Problémy s kvalitou dat pro řízeni

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?

Problémy s kvalitou dat pro řízeni

- 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

Anonimizace

- Zajistit, aby se nemohla zpětně identifikovat z info osoba, ke které data/info patří
- V plné míře obtížné
- Hlavní zádrhel – jak propojit k sobě patřící data pocházející z různých zdrojů a pořizovaných v různé době

Data a informace

- Hlídat výstupy aplikací generujících informace
  - Zda neprozrazují hlídaná data (info o jednom subjektu)
  - Závazek uživatelů, že nezneužijí takto kompromitované informace
- Aplikace na prověřeném serveru
- Výstupy logovat, kontrola výše uvedených závazků

Musí to ale být stanoveno zákonem

- *Současná situace vede ke kolosálním ztrátam nejen ve školství*

Kvalita dat musí být zohledňována ve specifikacích

- Kvalita dat může podstatně omezovat to, co je možné
  - Jednotlivé situace nejsou zdaleka zjevné.
  - Mnohé dimense kvality dat se opomíjejí.

Odpor proti kvalitě dat a informací

- Horší podniky nemusí mít zájem o zveřejňování informací, které odhalují jejich horší kvalitu, mohou být i jibné postranní úmysly
- Příklad sledování úspěšnosti absolventů škol. Principy:
  - Veřejný systém
  - Každý svoje kriteria hodnocení
  - Pro všechny školy

Špatně nastavená pravidla ochrany osobních dat - úzké místo veřejných informačních systémů

Brutální metody ochrany (osobních) dat mají chránit základní lidská práva

Dosahují ale opaku.

Ohrožují budoucnost IT a nejen jich

Ničení osobních dat jako ochrana před Velkým bratrem

Prý nutné pro splnění zásad Deklarace základních lidských práv a svobod , především práva na soukromí

- Data se de facto smí bez explicitního souhlasu dotčených osob používat a shromažďovat pouze k účelům, pro které byla pořízena a to jen pověřenými institucemi
- Každá data nevyhovující této podmínce musí být zničena, i třeba existují jen proto, že se objevil nový účel důležitý pro dotčený subjekt
- To nazveme **brutální proces ochrany dat (BPOD)**

Výchozí mlčky činěné předpoklady, vlastně předsudky

1. BPOD jsou v souhlasu s Deklarací základních lidských práv a jsou jejím důsledkem
2. BPOD umožňují efektivně chránit osobní data,
   - podstatně omezí počet případů, kdy mohou moje osobní data uniknout
3. BPOD nemají zásadní negativní sociální, celospolečenské a ekonomické efekty a nemají ani negativní dopady na informatiku
   – Předpokládá se tedy, že škody, ke kterým by došlo kompromitováním osobních dat pokud by se BPOD nepoužívala, jsou podstatně závažnější než důsledky nedostupnosti *zveřejnitelných* informací vypočitatelných z osobních dat + náklady na BPOD

Žádný z těchto předpokladů neplatí!!

Brutální procesy ochrany dat nezlepšují podstatně ochranu dat

- Pro každého je důležité, aby jeho osobní data nepřišla (neunikala) do nežádoucích rukou
  - jako osobě je mi jedno jakým způsobem a za jakým účelem.
- *Existuje ale mnoho kanálů úniku osobních dat a to BPOD nezmění!!*
- Některé existují ze zákona!!!!

Kanály úniků dat, některé je obtížné jiné nemožné uzavřít

- Mnohé údaje jsou veřejné ze zákona (obchodní rejstříky, registry nemovitostí, ..) a mnohé se z nich dá zjistit, jiné nejsou dostatečně zabezpečeny
- Některá data pacienta jsou např. pro léčbu natolik potřebná, že lékař považuje za správné je i přes zákazy využívat (jinak poruší Hippokratovu přísahu, de facto i zákon)
  - To oslabuje celý systém ochrany dat (legislativní disciplinu)
  - Ukazuje to, že není vše v pořádku

Kanály úniků dat, některé je obtížné jiné nemožné uzavřít

- Registry a rejstříky (katastrální, obchodní, občanů,spolků, …)
- Mobilní telefony
- Webové služby
- Sociální software
- Serverové stanice, cloudy (DATA JSOU LECKDE)
- Finanční instituce
- Zdravotní instituce
- Obchodování na webu (často partneři nejsou dostatečně profesionální a opatrní, někdy ani nemohou být)
- Atd. (špionážní satelity)

BPOD ohrožuje základní lidská práva, např. právo na život a na dobrou zdravotní péči

- Příklad zákazu SOA systému na online monitorování výdeje léků jako prevence výroby pervitinu
  - Blokoval se nadměrný výdej léků s pseudoefedrinem jedné osobě za krátkou dobu jako prevence výroby Pervitinu
  - Výroba Pervitinu skutečně významně klesla
  - Systém byl zakázán ÚOOÚ, neboť používal zdravotní data jednotlivých osob (léky, které používají)
- Ponecháváme stranou podezření, že někteří zúčastnění s takovým výsledkem předem počítali

BPOD ohrožuje základní lidská práva, např. právo na život a na dobrou zdravotní péči

Důsledky:

- Výroba Pervitinu se zase rozjela
  - Tragédie narkomanů a jejich rodin
  - Posílení podsvětí
  - Snížení prestiže státu u občanů

BPOD ohrožuje základní lidská práva, např. právo na život a na dobrou zdravotní péči

Důsledky 2:

Ztráta budoucích příležitostí:

- Nelze pomýšlet na on-line prevenci chybných medikací (ohrožení životů a zdraví),
  - To způsobuje ztráty životů na úrovni ztrát životů v dopravě (více než tisíc ročně),
    - v USA jsou kvalifikované odhady na úrovní cca 50000 ročně, takže u nás nějaké dva tisíce ročně, jistě existují kvalitnější odhady, základní zjištění platí a dá se použít i ve veřejných debatách.
  - Prevence chybných medikací by to mohla podstatně omezit počet vážných poškození zdraví.
    - V USA se odhaduje na cca 1,2 mil. ročně, takže u nás tak asi 50000 ročně. Počet postižených jde tedy do statisíců

BPOD ohrožuje základní lidská práva, např. právo na život a na dobrou zdravotní péči

Důsledky 3:

Ztráta budoucích příležitostí

- Zhoršení podmínek zdravotnického výzkumu a kvality reakce na epidemie,
- Blokování optimalizace systému zdravotních pojišťoven,
- Blokování účinné kontroly účinků léků, optimalizace léčby.
  - Pár miliard by to hodilo.
- Objev cest šíření cholery analýzou osobních dat provedený londýnským lékařem kolem r. 1850 by dnes byl nezákonný

BPOD ohrožuje základní lidská práva, např. právo na život a na dobrou zdravotní péči

- Zákaz platí i pro využívání dat zdravotních pojišťoven akreditovanými pracovišti
  - To už je naprostá zhovadilost
- Pro státní správu má tedy de facto přednost ochrana dat před ochranou životů a zdraví
  - Existují pověsti,že některé instituce se k tomu oficiálně hlásí
- Mělo by to být veřejnosti známo včetně hlavních důsledků!!!

Skryté omezování práva na informace a tedy na vzdělání

- Chybí nezávislý systém evaluace kvality škol a vzdělávání podle kriterií hodnotitele, např. rodiče
- Proto je obtížné vynucovat kvalitu výuky a správně volit směr studia a školu, není dohled nad efekty didaktických modernizací,
  - Stížnosti u nás i v USA (nedávno Obama)
  - Je dost indikací, že se kvalita vzdělání snižuje (STEM), ale je obtížné vyvolat změnu
  - Propad českých VŠ studentů v mezinárodních soutěžích

Data a informace

- *Samozřejmě není zájem horších škol situaci změnit*
  - *Jsou to dobré tunýlky*
- *Politici se snaží vyhovět voličům, i když ti nemusí chtít to nejlepší pro své děti (matematika)*
- *Může být zájem o to, aby lidé nebyli příliš vzdělaní*
  - *Spíše je to nedohlédnutí následků*

Kvalita dat a formulace požadavků na informační systémy

- To, co a jak můžeme uskutečnit je limitováno kvalitou dat více, než jsme ochotni připustit.
- Příklady:
  - Prostředky pro řízení projektů
    - Efekt líného studenta
    - Boj se zpevňováním norem
  - Řízení výrobních procesů
    - Data mohou být drahá nebo nutně neúplna (Franta se včera opil)

Možné řešení
Použít aparát analýzy rizik

- Sledovat všechna možná rizika, a všechna práva
  - Sledovat cenu prevence, porovnávat ji s očekávanými efekty
  - Zahrnovat i cenu vyvolaných rizik
- Prokazovat, že opatření skutečně dosahuje proklamované cíle (snižuje významně pravděpodobnost úniku a zneužití dat)
- Mělo by být kontrolovatelné

Kvalita dat a formulace požadavků, SW řízení projektů

- Metoda kritické cesty, MSProject
- 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 a data nadhodnocuje.
  - 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).

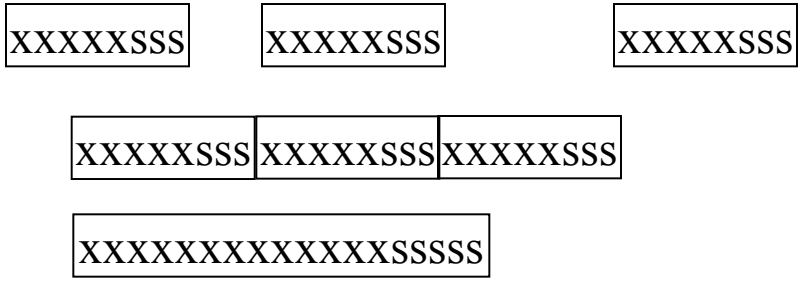
Důvody skluzů

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 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 zpevňování norem)
- Nestihne-li, nedá se nic dělat

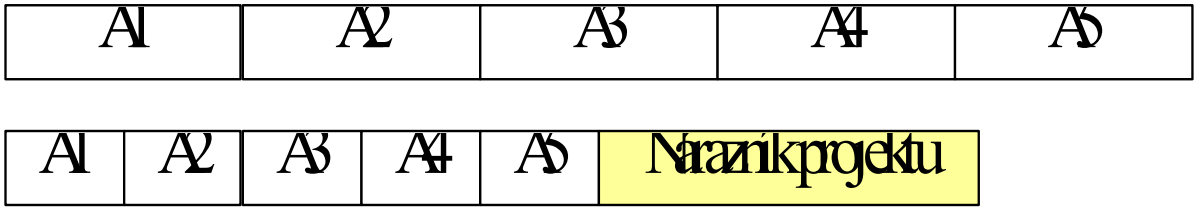
Kritický řetězec – doba řešení je součtem nezávislých n.v.



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

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

Přípojné buffery



$$\textit{Délka nárazníku} \approx \Sigma\,(3\sqrt{\Sigma\sigma_i^2})$$

Řešení

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

Řešení, kritický řetězec

- Kritický řetěz funguje jen když jsou řešitelé ochotni odhalit reservy 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

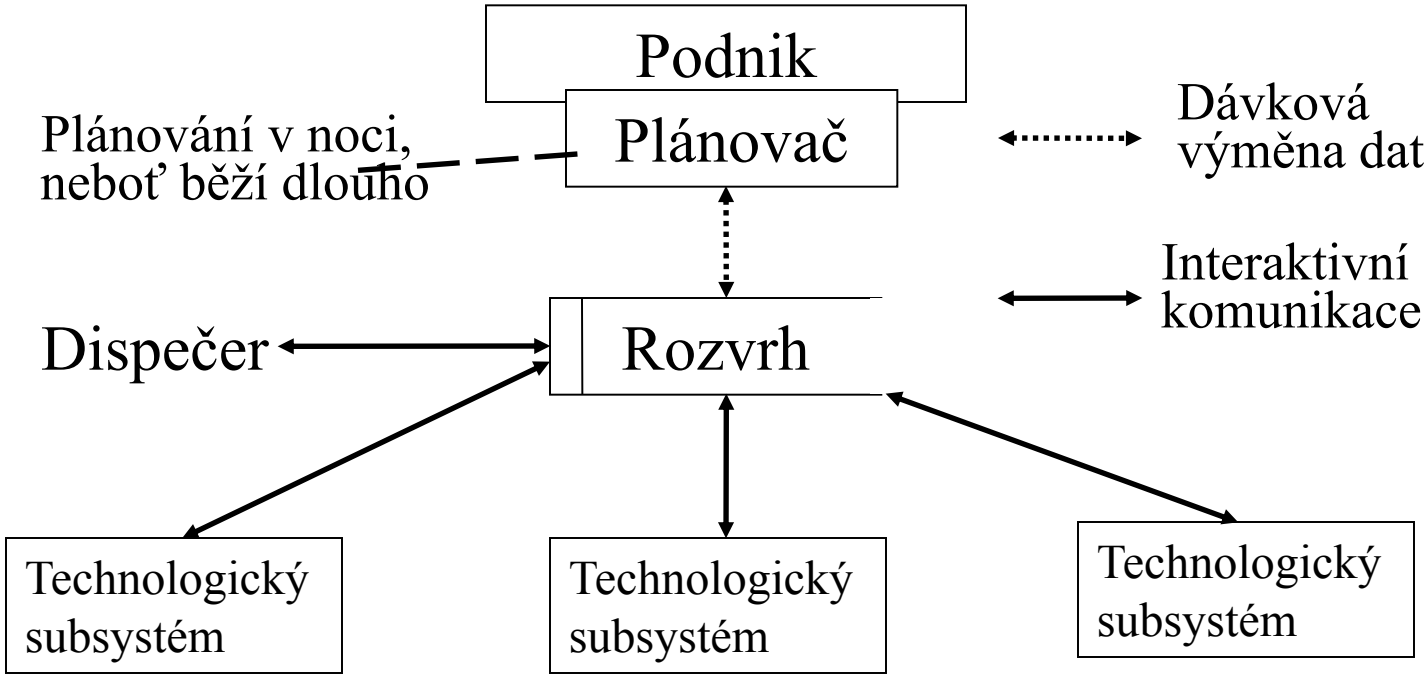
- Každý začne pracovat hned, jak je jeho předchůdce hotov. To ale znamená, že je nutné nějak vyloučit efekt multitaskingu, To se řeší tak, že se dané činnosti postupně stále přesněji oznamuje, kdy bude třeba začít pracovat na daném projektu (činnost je tedy spravována jako autonomní služba). (Zpřesňování 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 chybějících a pomalu aktualizovaných dat

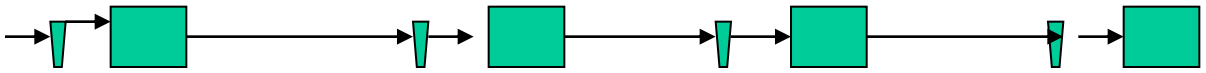
- Rozvrhování na úrovni podniku nemůže mít všechna data v dobré kvalitě
  - Cena sběru, některá data se nesbírají
  - Jsou nedostupná (taktilní, zkušenostní, blokované znalosti)
  - Nedostatečně přesná
- Algoritmus rozvrhování musí být pomalý (exponenciální složitost, nutné jsou dohody s lidmi) a nepřesný v důsledku nekvality dat – použít úložiště dat (nutné 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



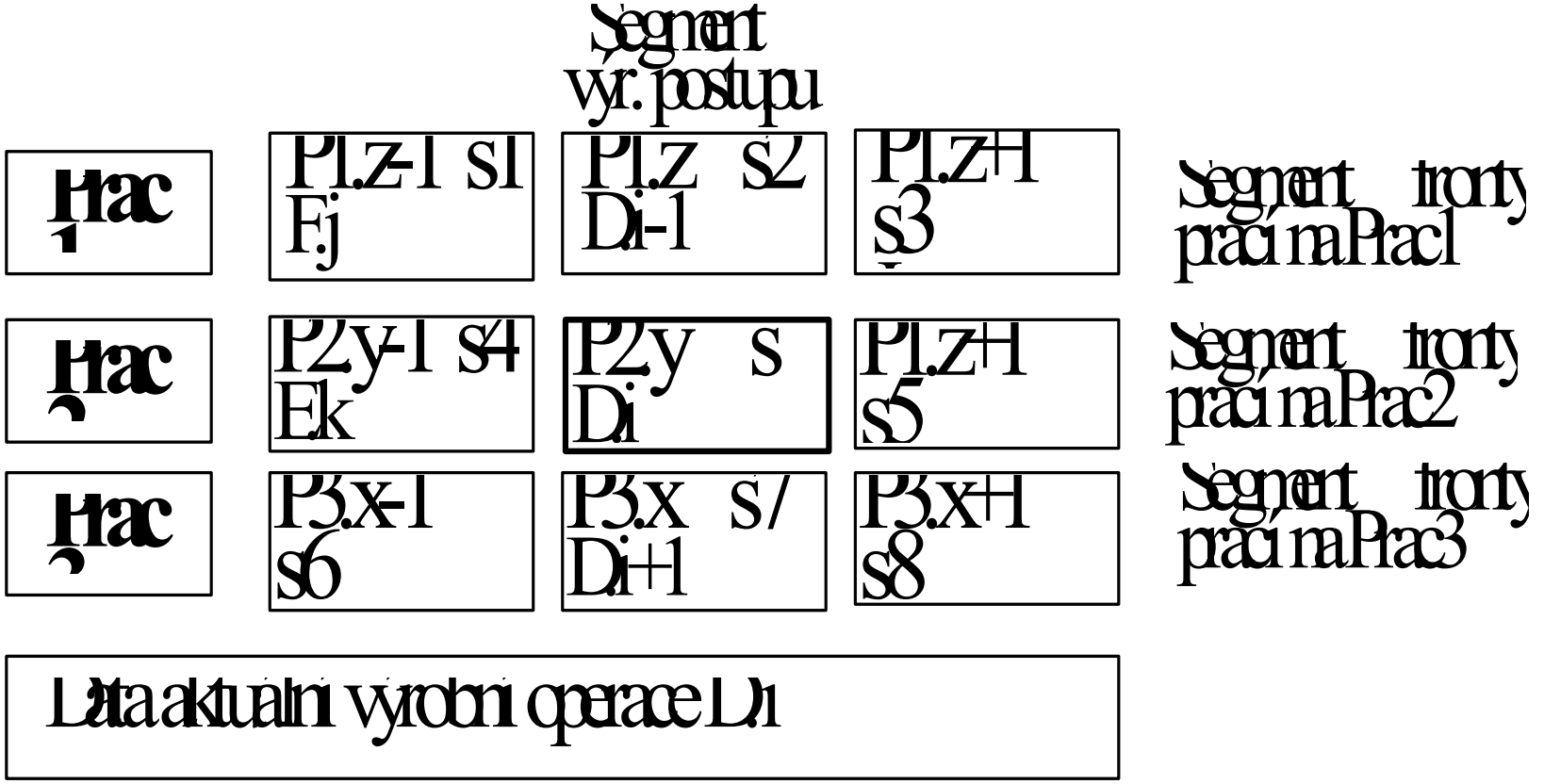
Řízení na buffery



Princip

Mistr sleduje bufery a hledá pro dané pracoviště práci, když se jeho bufer nepříjemně zkracuje (řízení na průšvih)

Nutné pro výroby s krátkými seriemi a velkou variabilitou prací


Segment
výr. postupu
Prac 1
P1.Z-1 S1
Fj
P1.Z S2
Di-1
P1.Z+1
S3
Segment fronty
prací na Prac1
Prac 2
P2.y-1 S4
Ek
P2.y S
Di
P1.Z+1
S5
Segment fronty
prací na Prac2
Prac 3
P3.x-1
S6
P3.x S7
Di+1
P3.x+1
S8
Segment fronty
prací na Prac3
Data aktuální výrobní operace Di

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)
- Potřeba využít inteligenci lidí jako součásti procesů

Problémy s kvalitou dat pro řízeni

- 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?

Problémy s kvalitou dat pro řízeni

- 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. ?UML?
- Může podstatně ovlivnit použitelnost sématického webu.
- Srozumitelnost a deklarativnost dat na rozhraních je klíčovou podmínkou použitelnosti business procesů. A co pak objektová orientace?

Závěr 3

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

Pravidla hry podrobněji

1. Především změníme způsob plánování prací. Nebudeme stanovovat, kdy se přesně na jednotlivých etapách začne pracovat a kdy práce skončí. Místo toho se stanoví, jak dlouho bude asi řešení etapy trvat (např. odhad střední doby práce nebo mediánu, často se volí polovina odhadu *H* horní hranice kontigenčního intervalu), a kdy se asi na ní bude moci začít pracovat.

Pravidla hry podrobněji

1. Práce na etapě se zahájí co nejdříve od okamžiku, kdy je to možné práci zahájit. Aby tomu bylo možné vyhovět, je postupně zpřesňován odhad okamžiku, kdy bude možné začít na etapě pracovat. K tomu je nutné mít průběžné informace o stavu řešení předcházejících etap. Zkušenost ukazuje, že tento postup skutečně umožňuje, aby řešitelé zorganizovali práci tak, aby mohli začít na projektu pracovat hned, jak je to možné.

Pravidla hry podrobněji

1. Od okamžiku zahájení prací se pracuje pouze na úkolech spojených s řešením etapy a pracuje se s maximální intenzitou (to vylučuje efekt líného studenta a multitaskingu).
2. Řešitelé dostatečně často předem hlásí, kdy asi budou hotovi a práci odevzdávají hned, jak jsou hotovi (to je nutné pro bod 1).

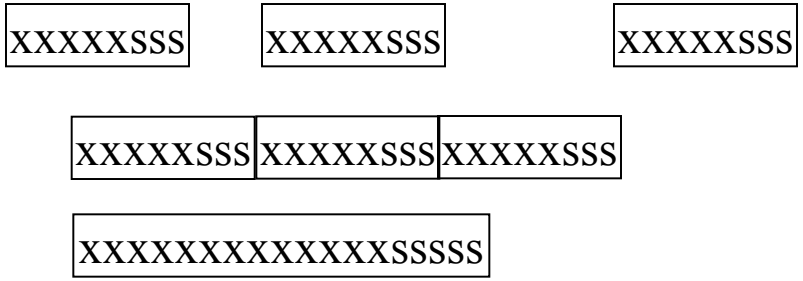
Pravidla hry podrobněji

- Těmto požadavkům lze vyhovět jen tehdy, kdy budou řešitelé ochotni pracovat naplno a odhalovat své rezervy. Musí být proto zainteresováni na úspěchu řešení projektu a musí mít také jistotu, že se proti nim nepoužije procedura zpevňování norem.

Pravidla hry podrobněji

- Vedení projektu musí naopak chápat, že jsou termíny kruté a že se často nesplní. Z nesplnění termínu by neměly být zpravidla vyvozovány žádné postihy. Všichni by měli mít prospěch ze zkrácení doby řešení a z prémií za včasné dokončeni projektu. Podstatnou roli tedy hrají psychologie a sociální aspekty fungování týmu. To je u IS standardní situace.
- Je dobré provádět analýzu dat, kdy a jak je kdo hotov s cílem odhalit lenochy

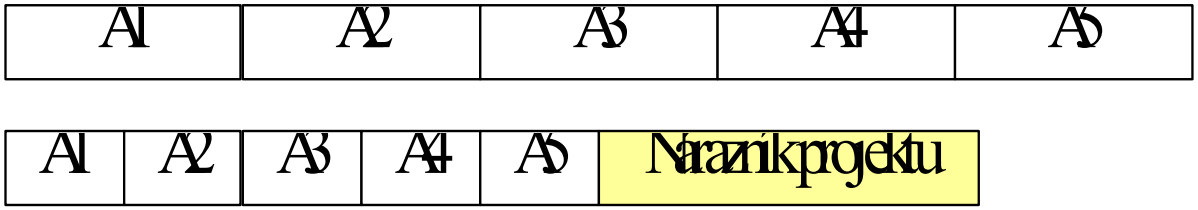
Kritický řetězec – doba řešení je součtem nezávislých n.v.



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

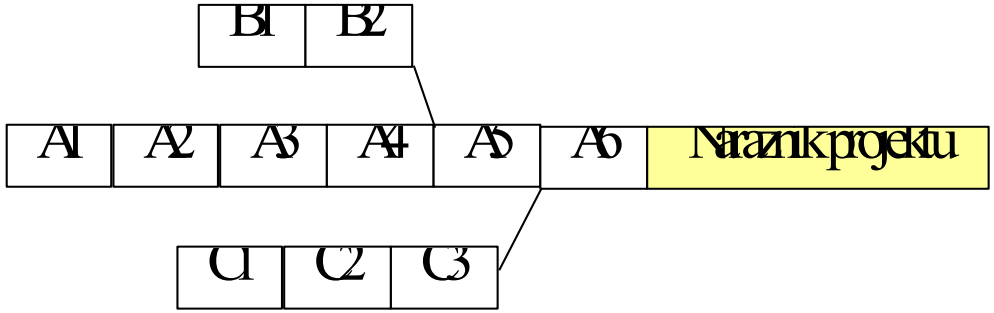
$$T \approx \Sigma\ (t_i + 3\sqrt{\Sigma\ \sigma_i^2})$$

Přípojné buffery

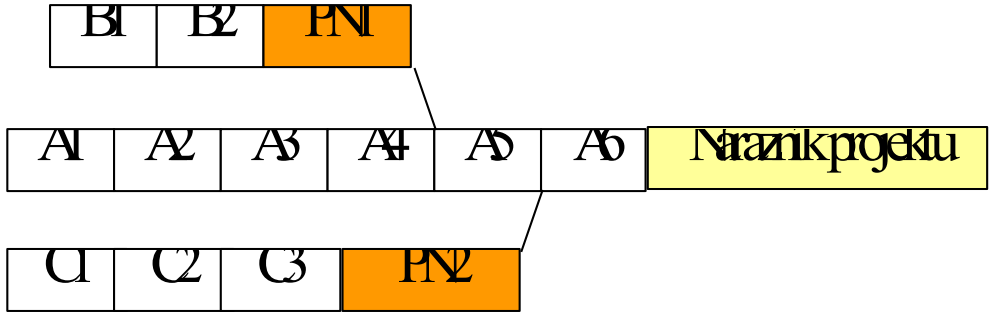


$$\textit{Délka nárazníku} \approx \Sigma\,(3\sqrt{\Sigma\sigma_i^2})$$

Přípojné buffery



Obr. KC2. Kritická cesta s nárazníkem projektu A1, .., A5 je kritická cesta



Obr. KC3. Projekt s přípojnými nárazníky PN1 a PN2

Práce s přípojnými buffery

- Předpokládejme, že návaznost činností v projektu tvoří více lineárních úseků, které se postupně spojují. V tom případě najdeme kritickou cestu. Připojíme za ní nárazník projektu jehož velikost je dána rezervami činností na kritické cestě. Činnosti mimo kritickou cesto zobrazíme jako větve stromu (obr. KC2). Aby činnosti mimo kritickou cestu a projekt se dal lépe řídit je žádoucí doplnit nárazníky délky vypočtené výše uvedeným způsobem pro každou postranní větev. Tyto nárazníky nazveme přípojné nárazníky
- Při řízení projektu se sledují pro každou větev výše uvedeným způsobem přípojné nárazníky a nárazník projektu. Pokud přípojný nárazník nestačí (je vyčerpán) zkrátí se i nárazník projektu.

Soutěž o zdroj X

- Klasická metoda kritické cesty nedostatečně zvažuje případ, kdy je nějaká činnost X prováděna na více větvích, nebo dokonce ve více projektech (příkladem mohou být kontrolní nebo dokumentační činnosti). Řešení tohoto problému je poměrně komplikované, dobře pracuje následující přiblížení.

Soutěž o zdroj X

a) Činnosti X se považují za činnosti na kritické cestě a proto přispívají standardním způsobem ke zvětšení projektového nárazníku (tato cesta se nazývá kritický řetěz).

b) Před každou činností X se vytváří přípojný nárazník.

c) Pro činnost X je možné vytvořit frontu prací obsluhovanou v závislosti na termínech navazujících etap v projektu.

d) Pokud je X úzké místo celé firmy (do značné míry určují výkon firmy) dostávají přednost požadavky těch projektů, pro které má zlomek (výnos projektu)/(doba vytížení X) maximální hodnotu.

Komplikovanější řešení používá i plánování kritického zdroje (podle teorie omezení bývá jen jeden).

Hodnocení

- Mnohé nedořešeno (obecné acyklické grafy)
- Osvědčuje se podle dostupných zpráv
- Chtělo by to asi lepší statistické zpracování obecnějších případů

Hodnocení

- Krásný příklad, jak řešení závisí na kvalitě dat a také na tom, že i pak je řešení závislé na na dobrých vztazích v podniku, jeho kultuře a morálce.
- Považuji-li lidi za onuce, nemohu čekat dobré výsledky
- Psychologický kapitál může být zatraceně významný
- Dobrý vztah k lidem není věc dobročinnosti ale chladného kalkulu

Dimense kvality informací

Quality Categories	Information Quality Dimensions
Intrinsic IQ	Accuracy, objectivity, believability, reputation
Contextual IQ	Relevancy, value-added, timeliness, completeness, amount of information
Representational IQ	Interpretability, ease of understanding, concise representation, consistency
Accessibility IQ	Access, security

Table 2 Information Quality Categories and Dimensions (Source: Wang, Strong [10])

Dimense kvality informací

Sequence of examination				Examples of attributes of data/information quality		
				Direct Attributes		*Indirect* Attributes
	▼	*Primary* Attributes	*Mandatory* Attributes	**Interpretable**		Legible, user trained, untrained, educated
				Significantly relevant		Concise, current, admissible, secure, appropriate amount
				Critically timely		Obtainable, accessible, style and mode of decision making, individual or collective
			Desirable Attributes	**Critically credible**		Believable, reputable, decision maker's traits: risk averse, passive, hesitant, cautious, prudent, motivated, jumpy
				Acceptably complete (for totality of factors)		
	Irrelevant	*Secondary* Attr.		Economically	**Timely**	Frequency, how much in advance
					Unbiased	Sampling, observation points,
					Accurate (error-free)	Mapping (complete, unambiguous, meaningful, and correct), granularity, age
					Precise	Number of significant digits, dots
					Easy to use	How summarized, detail, text, graph, diagram, picture, media, clarity, order, consistent, homogene, understandable, natural, efficiently encoded

Table 4. Example of a hierarchical result-oriented taxonomy of data or information quality attributes in economical sequence of their examination

Anonimizace

- Zajistit, aby se nemohla zpětně identifikovat z info osoba, ke které data/info patří
- V plné míře obtížné
- Hlavní problém – jak propojit k sobě patřící data pocházející z různých zdrojů a pořizovaných v různé době a neprozrazovat identitu osob

Data a informace

- Hlídat výstupy aplikací generujících informace
  - Zda neprozrazují hlídaná data (info o jednom subjektu)
  - Závazek uživatelů, že nezneužijí takto kompromitované informace
- Aplikace na prověřeném serveru
- Výstupy logovat, kontrola výše uvedených závazků

Musí to ale být stanoveno zákonem

- *Současná situace vede ke kolosálním ztrátam nejen ve školství*

Kvalita dat musí být zohledňována ve specifikacích

- Kvalita dat může podstatně omezovat to, co je možné
  - Jednotlivé situace nejsou zdaleka zjevné.
  - Mnohé dimense kvality dat se opomíjejí.

Odpor proti kvalitě dat a informací

- Horší podniky nemusí mít zájem o zveřejňování informací, které odhalují jejich horší kvalitu, mohou být i jibné postranní úmysly
- Příklad sledování úspěšnosti absolventů škol. Principy:
  - Veřejný systém
  - Každý svoje kriteria hodnocení
  - Pro všechny školy

Data a informace

- *Samozřejmě není zájem horších škol situaci změnit*
  - *Jsou to dobré tunýlky*
  - *Lenost a pohodlnost, i se strany rodičů*
- *Politici se snaží vyhovět voličům, i když ti nemusí chtít to nejlepší pro své děti (matematika)*
- *Může být zájem o to, aby lidé nebyli příliš vzdělaní*
  - *Spíše je to nedohlédnutí následků*

Jaký je hlavní úkol škol

- Vytvořit pracovní návyky
  - Předat dovednost, to lze jen tréningem (drilem)
  - Umět získávat znalosti a umět je používat